Development of an Instrument to Measure Staff-Reported Resident-to-Resident Elder Mistreatment (R-REM) Using Item Response Theory and Other Latent Variable Models

Jeanne A. Teresi, EdD, PhD, Katja Ocepek-Welikson, MPhil, Mildred Ramirez, PhD, Joseph P. Eimicke, MS, Stephanie Silver, MPH, Kimberly Van Haitsma, PhD, Mark S. Lachs, MD, MPH, PhD, and Karl A. Pillemer, PhD

1Research Division, Hebrew Home at Riverdale, Bronx, New York.
2Stroud Center, Columbia University, and New York State Psychiatric Institute, New York.
3Division of Geriatrics and Palliative Medicine, Weill Cornell Medical College, New York.
4Abramson Center for Jewish Life, Polisher Research Institute, North Wales, Pennsylvania.
5Division of Geriatrics and Palliative Medicine and the Center for Aging Research and Clinical Care, Weill Cornell Medical College, New York.
6Department of Human Development, Cornell University, Ithaca, New York.

*Address correspondence to Mildred Ramirez, PhD, Research Division, Hebrew Home at Riverdale, 5901 Palisade Avenue, Riverdale, NY 10471.
E-mail: milramirez@aol.com

Received September 24, 2012; Accepted January 2, 2013
Decision Editor: Rachel Pruchno, PhD

Objectives: Although numerous studies have measured behaviors among individuals in congregate settings, few have focused on resident-to-resident elder mistreatment (R-REM). To our knowledge, there is no psychometrically developed measure of R-REM extant. The quantitative development of a measure of staff-reported R-REM is described. Methods: The design was a prevalent cohort study of residents of 5 long-term care facilities. The primary certified nursing assistant was interviewed about R-REM. Advanced measurement methods were used to develop a measure of R-REM. Results: The loadings on the general factor for the final 11-item scale were greater than those on the group factor except for the item “other physical behavior” (0.63 vs. 0.74), suggesting essential unidimensionality. Although the bifactor model fit was slightly better than that of the unidimensional model, the difference was trivial (bifactor comparative fit index [CFI] = 0.997, root mean square error of approximation [RMSEA] = 0.013, unidimensional CFI = 0.979, and RMSEA = 0.030). However, modest support was provided for use of verbal and physical subscales. The explained common variance statistics were 0.76 for the bifactor model compared with 0.63 for the unidimensional model. Discussion: The development of this R-REM measure will help to advance the measurement and ultimately evaluation of interventions associated with this important and under recognized problem facing residents in long-term care settings.

Key Words: Elder abuse, Mistreatment, Congregate care, Item response theory, Latent variable models, Scale development

Although numerous studies have identified, classified, and measured behaviors among individuals in congregate settings and among those with cognitive impairment (Burgio et al.,
Behaviors directed by residents against other residents are potentially serious because they can be sudden and unexpected, leaving the recipients fearful in anticipation of the next occurrence. The prevalence of R-REM has not been demonstrated fully; however, Lachs, Bachman, Williams, and O’Leary (2007) found that almost 6% of community residents had police contact after being placed in a long-term care institution, the majority of which were due to R-REM.

A recent study of combative behaviors of residents (Morgan et al., 2012), although not focused on R-REM, used diaries and questionnaires completed by frontline staff to document incidents of aggression directed toward them. The majority of staff reported occurrence of incidents within the past month, providing evidence regarding the ubiquity of such aggressive behaviors. Although there are many factors that contribute to resident-to-resident aggression, chief among them is the commingling of residents with a mixture of cognitive impairment, comorbidity, and differing lifestyles and personalities in small places (Teresi, Holmes, & Monaco, 1993). For example, in one of the few studies of R-REM extant, Shinoda-Tagawa and colleagues (2004) found that cognitive impairment, aggressive behavior, and physical independence were associated with physical abuse of recipients. Measurement in the study by Shinoda-Tagawa and colleagues (2004) was from reports to the Department of Health and abstraction of police reports. Most recently, Castle (2012), building on the conceptual work of Lachs and Pillmer (2004) and Stiles, Koren, and Walsh (2002), assessed resident-to-resident abuse using a mailed questionnaire completed by certified nursing assistants (CNAs), with 35 items measuring five domains: verbal, physical, psychological, sexual abuse, and material exploitation (taking other peoples belongings). Although the study was large, with facilities sampled from 10 states, the authors discuss that more refined measures focused on specific residents rather than aggregated reports are needed to determine prevalence and better understand the mechanisms of R-REM.

The first study of abuse of patients in nursing homes (Pillemer & Moore, 1989) resulted in a taxonomy of abuse, based on earlier work in the community (Pillemer & Finkelhor, 1988) using an instrument designed to measure forms of family violence, on which the instrument derived for this study was partially based. The methodology has been replicated in several other countries and has produced remarkably stable estimates of self-reported community elder abuse (Lachs & Pillmer, 2004). There is no “gold standard” in violence research because researchers or anyone outside the violent dyad rarely witnesses actual episodes. Self-report is widely accepted as the primary methodology for case finding in violence research, and this method has been used in this study. Given the high prevalence of cognitive impairment in long-term care settings, staff reports are critical in the assessment of R-REM.

Aims

The purpose of this article is to describe the quantitative development of a measure of staff-reported R-REM (R-REM-staff version; R-REM-S). The qualitative development is described elsewhere (Ramirez et al., in press).

Conceptual Model

The conceptual model is derived from earlier work by Pillmer and Moore (1989) and by qualitative research using focus groups (Rosen et al., 2008). In the latter study, R-REM was categorized as physical, verbal, and sexual and 35 types were identified within these classifications. Later, Pillmer and colleagues (2011) characterized R-REM in terms of typologies, identifying five categories as follows: invasion of privacy, roommate altercations, hostile interpersonal interactions, unprovoked actions, and inappropriate sexual behaviors. Qualitative work (Ramirez et al., in press) identified four broad domains of R-REM: verbal, physical, sexual, and other (e.g., wandering into others’ rooms). These domains form the basis for the measure presented here. For the purpose of this study, resident-to-resident mistreatment is defined as negative and aggressive physical, sexual, or verbal interactions between residents in long-term care settings, which in a community setting would likely be considered as unwelcome.

The general approach taken was derived from a long-standing and well-established
measurement strategy in the field of interpersonal aggression. The best known of the instruments based on this strategy is the Conflict Tactics Scale (CTS; Straus, Hamby, Boney-McCoy, & Sugarman, 1996). Measures based on the CTS model have been successfully used with nursing home staff (Pillemer & Moore, 1989). The CTS conforms to a major finding regarding effective measurement of mistreatment, that is, violence measures must be specific. Behaviorally specific questions have a number of advantages: They require fewer definitions of terminology (as would a general question such as “Have you ever been abused?”), they are less subject to a respondent’s interpretation of terms, and they are less dependent on educational level.

Most important, behavioral specificity reduces the effects of potential personal and cultural differences in interpretation of terms such as “abuse” or “neglect” (Acierno, 2003). A further strength of the CTS approach is separating the focus on behaviors from outcomes. The approach is consistent with legal definitions, which negatively sanction assault regardless of whether it results in serious injury. R-REM may also be serious but does not cause obvious physical harm. Shinoda-Tagawa and colleagues (2004) included only individuals who had suffered a documented injury in their study. This rule is likely to exclude individuals who may be psychologically distressed by an incident or who may have been injured in ways that are not obvious (Herbert & Bradshaw, 2004). Thus, the present measure moves beyond the focus only on R-REM that has resulted in injury and includes a wide range of actions, regardless of outcome.

One central feature of R-REM may separate it from nearly every other form of interpersonal violence—it occurs in an environment that is both a residence as well as a health care facility. Accordingly, many health care providers are potentially “stationed” to detect R-REM either directly or indirectly. To not make systematic use of these observers would be to lose an opportunity for more comprehensive case finding and description. Thus, this study used staff reports via a staff version of the R-REM measure, after brief in-service training that is part of customary periodic facility education (e.g., infection control).

Although few systematic studies of staff reporting of R-REM exist, it is noted that nursing home staff have been “violence and conflict reporters” in other research studies not involving R-REM. Remarkably, nursing home staff members have actually reported their own acts of mistreatment (anonymously) toward residents as well as witnessed acts of abuse by other staff (Pillemer & Moore, 1989). These studies would suggest that nursing home staff could be reliable informants for detecting and describing R-REM, particularly because it is not as personally stigmatizing as self-promulgated abuse.

**Methods**

**Design**

The design was a prevalent cohort study of all residents who consented or were consented by designated proxies (families or legal guardians). The study was approved by Institutional Review Board, protocol 0803009718.

**Sample**

*Selection of facilities.*—Using a pseudo-random number generator procedure, 6 nursing homes were selected from among the population of 21 nursing homes with 250 or more beds in two New York regions. Large facilities were selected to ensure a sufficient number of units to permit randomization within facilities. Two facilities with severe survey deficiencies were excluded. Agreement to participate was obtained from five of the six facilities. The final sample represented 24% of large facilities (250 plus beds) in the two regions.

**Procedures**

The CNA who was most involved with the resident’s care (usually the primary care CNA) was interviewed by research staff using a structured interview schedule via a computerized assisted interview.

**Measure**

In addition to the CTS, the staff R-REM instrument was developed after extensive modifications to some items from the Cohen-Mansfield Agitation Instrument (CMAI; Cohen-Mansfield et al., 1989), an instrument that measures behavioral disruption in nursing home residents. The qualitative work in modification and translation is described elsewhere (Ramirez et al., in press). After focus groups with nursing home staff, more R-REM items were added so as to create a new instrument: the resident-to-resident elder mistreatment-staff version.
(R-REM-S). The items range from those low on aggressiveness (such as screaming at another resident) to increasingly aggressive tactics involving physical violence.

In following the tradition of elder mistreatment reporting in community studies, both a global dimension and subtypes of R-REM (e.g., physical, verbal, and so on) were posited. The original item pool contained 22 items across four domains: verbal (five items, e.g., “cursing,” “intimidation,” “ethnic slurs,” and so on), physical (seven items, e.g., “hitting,” “kicking,” “scratching,” and so on), sexual (three items, e.g., “saying sexual things,” “inappropriate touching,” and so on), and others (seven items, e.g., “unwanted help,” “threatening gestures,” “wandering,” and so on).

The R-REM-S instrument was designed to collect data on staff recollection of resident-to-resident mistreatment during the past 2-week period. A 2-week period was selected because this time period is often used in psychiatric and behavioral measures addressing underlying latent constructs that may be rare. A list of all potential R-REM behaviors is provided and the number of distinct incidents involving these behaviors is requested. For each incident (up to five), the behaviors involved, the location (e.g., dining area, hallway, resident’s room), and the time (e.g., morning, noon meal, afternoon) it occurred are recorded. In addition, a description of other participant(s) (sex and relationship) and who started it are also recorded. The CNA reports about whether he or she witnessed the incident, and if so, what he or she did about this (e.g., separated residents, redirected residents). Excluding incidents that occurred within the past 2 weeks, the inquiry then focused on the two most serious occurrences within the past year. Finally, the staff were asked to describe the most serious incident, if any that was directed toward staff.

**Statistical Procedures**

Increasingly, advanced psychometric methods are being used to assess gerontological and health-related measures (Fang et al., 2012; Reeve et al., 2007; Teresi, Stewart, Morales, & Stahl, 2006). According to the conceptual model discussed previously, the items are effect indicators in that the underlying latent variable R-REM gives rise to the indicators and accounts for the variance in the item set. Further, it is posited that the measure is essentially unidimensional with one single underlying attribute but that additional variance arises due to factors that may reflect R-REM or nuisance variables. First, traditional methods of examining dimensionality were performed. This is the method similar to that described in Asparouhov and Muthén (2009), in which a merged exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are performed fitting a unidimensional model with polychoric correlations. This method uses the exploratory analyses to permit cross-loadings on variables, followed by a confirmatory model with loadings specified on only one factor. The exploratory analyses used principal components estimation and examined tests of scree with cross-loadings permitted. This was followed by the confirmatory analyses of the unidimensional model. Additionally, a two-factor model was tested against a unidimensional model to determine if the fit improved (results not reported).

However, to further investigate unidimensionality, a bifactor model was tested. The bifactor model assumes that a single general trait explains most of the common variance but that group traits explain additional common variance for item subsets (Reise, Moore, & Haviland, 2010). Thus, a general and two or more group factors are posited. Following the methods of Reise and colleagues (2010) and that of Revelle and Zinbarg (2009), a Schmid-Leiman (S-L; Schmid & Leiman, 1957) transformation (using the “psych” R package; Rizopoulus, 2006, 2009) was performed to find an alternative set of group factors for the restricted bifactor (Gibbons et al., 2007) model without cross-loadings. A three-factor model was specified with uncorrelated factors. The S-L procedure is as follows (Jennrich & Bentler, 2011; Reise et al., 2010). First, an oblique rotation factor analysis using the covariance matrix is performed to produce factor loadings, correlations, and unique variances. Then, a single factor analysis of the correlations is performed to extract second-order factors, followed by the S-L orthogonalization of the oblique results. The purpose is to obtain loadings for each item on uncorrelated general and group factors. However, it is noted that the S-L solution is used to guide the final bifactor estimation of loadings using MPLUS because the loadings produced by the S-L solution have been found to be biased (Reise et al. [2010]; see Jennrich and Bentler [2012] for alternative bifactor methods). This procedure permits inspection of the final loadings on the general as contrasted with the group factors and for comparison with a single
factor solution in sensitivity analyses to determine if essential unidimensionality is achieved and if subscales such as physical and verbal mistreatment are of any value.

All items were specified to load on the general factor, and the loadings on the group factors were specified following the S–L solution shown in Supplementary Figure 2. The bifactor model was run with the group factors as described previously using MPLUS (version 6.1; Muthén & Muthén, 1998–2010). A robust parameter estimation procedure, based on a mean- and variance-adjusted weighted least squares procedure (WLSMV) contained in MPLUS was used.

Summary statistics from the various models are presented in Table 1. Factor loadings are presented along with communalities \((h^2)\) and general factor variance. Both eigenvalues and fit statistics (root mean square error of approximation [RMSEA] and comparative fit index [CFI]) were examined but were not central in decision making because of the desire to find alternative approaches to the use of fit statistics for evaluating model adequacy.

Tests of unidimensionality.—The explained common variance (ECV) provides information about whether the observed variance covariance matrix is close to unidimensionality (Sijtsma, 2009). The ECV can be estimated as the percent of observed variance explained. It is the ratio of the first eigenvalue to the sum of all eigenvalues extracted from a bifactor model analysis and was calculated as the eigenvalue for the general factor divided by the sum of the eigenvalues for the general and group factors from a bifactor model (Reise et al., 2010):

\[
ECV = \frac{\sum \lambda^2_{\text{Gen}}}{\sum \lambda^2_{\text{Gen}} + \sum \lambda^2_{G1} + \sum \lambda^2_{G2} \cdots \sum \lambda^2_{GK}}.
\]

Reise and colleagues also suggest examining the difference in the loadings between the unidimensional model and the general factor loadings \((\lambda_1)\) in the bifactor model as an indication of the degree of distortion that would occur by fitting a unidimensional model to data that are multidimensional.

Tests of reliability.—The methods for assessing reliability that have emerged from the psychometric literature as the preferred statistics include McDonald’s (1970, 1999) omega total and hierarchical \((\omega_t\) and \(\omega_h\)) derived from linear structural equation modeling (SEM) and those derived from nonlinear SEM. Under the single common factor model, reliability can be evaluated by decomposing the scale score into the sum of the item scores, and the contribution of the common term \((\lambda F)\) or communality. Known as McDonald’s \((1999)\) omega total \((\omega_t)\), this reliability estimate is based on the proportion of total common variance explained. Alternatively, this can be expressed in terms of the unique variance, where the unique variance contains both specific and error variance:

\[
\omega_t = 1 - \frac{\sum u^2}{V_x}.
\]

(Revelle and Zinbarg, 2009; formula 19). Omega hierarchical \((\omega_h; McDonald, 1970)\) is calculated as the sum of the squared loadings on the general factor divided by the total scale score variance \((V_x; Revelle & Zinbarg, 2009)\): For the one- or two-factor model, \(\omega_h\) is not meaningful because three factors are required for identification in the calculation of \(\omega_h\). However, omega total \((\omega_t)\) is interpretable for a unidimensional solution.

Item response theory parameter estimation.—Additionally, item response theory (IRT), applying the two-parameter logistic model (Lord & Novick, 1968) was used to evaluate the measure. IRT has been used to model gerontological data sets. See Teresi, Cross, & Golden, 1989; Teresi, Kleinman, & Ocepek-Welikson, 2000 for didactic explanations of the models in that context. The estimates for the discrimination and severity parameters \((a, b)\) were evaluated, the item and test information functions were graphed, and the reliability estimates were calculated for points along the dimension of the underlying construct, denoted as \(\theta\) (theta).

Tests of IRT model fit and local dependencies.—Various goodness of fit statistics are available to test IRT model fit (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006). Residual covariances can be examined in the context of violation of the assumption of local independence. The local dependency (LD) statistics (Chen & Thissen, 1997) compare the observed and expected frequencies associated with cross-tabulations of item pairs. Standardized \((z)\) scores are used for comparison across items with different response categories: \(z = \chi^2 - df / \sqrt{2}df\) (Thissen, 2011). If the observed covariation
<table>
<thead>
<tr>
<th>Item description</th>
<th>One factor, $\lambda$ (SE)$^a$</th>
<th>Three factor</th>
<th>Two factor</th>
<th>Three factor</th>
<th>Two factor</th>
<th>S–L bifactor solutions</th>
<th>MPLUS bifactor solutions</th>
<th>S–L-based model$^b$</th>
<th>Final model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G$\lambda$ (var.)</td>
<td>F1$\lambda$</td>
<td>F2$\lambda$</td>
<td>F3$\lambda$</td>
<td>$h^2$</td>
<td>G$\lambda$ (var.)</td>
<td>F1$\lambda$</td>
<td>F2$\lambda$</td>
<td>F3$\lambda$</td>
</tr>
<tr>
<td>Use bad words toward another resident</td>
<td>0.91 (0.02)</td>
<td>0.78 (0.73)</td>
<td>0.42</td>
<td>0.22</td>
<td>0.83</td>
<td>0.71 (0.61)</td>
<td>0.57</td>
<td>0.83</td>
<td>0.88 (0.05)</td>
</tr>
<tr>
<td>Scream at another resident</td>
<td>0.95 (0.02)</td>
<td>0.80 (0.69)</td>
<td>0.52</td>
<td>0.92</td>
<td>0.75 (0.62)</td>
<td>0.58</td>
<td>0.89</td>
<td>0.87 (0.07)</td>
<td>0.47 (0.14)</td>
</tr>
<tr>
<td>Try to scare, frighten, or threaten (another resident) with words</td>
<td>0.85 (0.03)</td>
<td>0.72 (0.65)</td>
<td>0.48</td>
<td>0.20</td>
<td>0.79</td>
<td>0.64 (0.49)</td>
<td>0.64</td>
<td>0.82</td>
<td>0.79 (0.08)</td>
</tr>
<tr>
<td>Boss around/tell another resident what to do</td>
<td>0.79 (0.03)</td>
<td>0.62 (0.50)</td>
<td>0.61</td>
<td>0.78</td>
<td>0.58 (0.55)</td>
<td>0.51</td>
<td>0.60</td>
<td>0.71 (0.08)</td>
<td>0.40 (0.16)</td>
</tr>
<tr>
<td>Hit another resident</td>
<td>0.81 (0.04)</td>
<td>0.68 (0.62)</td>
<td>0.52</td>
<td>0.75</td>
<td>0.71 (0.68)</td>
<td>0.47</td>
<td>0.75</td>
<td>0.74 (0.06)</td>
<td>0.49 (0.10)</td>
</tr>
<tr>
<td>Grab or yank another resident</td>
<td>0.75 (0.06)</td>
<td>0.62 (0.56)</td>
<td>0.54</td>
<td>0.68</td>
<td>0.66 (0.64)</td>
<td>0.50</td>
<td>0.69</td>
<td>0.62 (0.08)</td>
<td>0.58 (0.12)</td>
</tr>
<tr>
<td>Push or shove another resident</td>
<td>0.74 (0.06)</td>
<td>0.61 (0.69)</td>
<td>0.23</td>
<td>0.55</td>
<td>0.63 (0.74)</td>
<td>0.25</td>
<td>0.54</td>
<td>0.71 (0.07)</td>
<td>0.34 (0.14)</td>
</tr>
<tr>
<td>Throw things at another resident</td>
<td>0.82 (0.05)</td>
<td>0.79 (0.64)</td>
<td>0.59</td>
<td>0.59</td>
<td>0.99 (0.72)</td>
<td>0.38</td>
<td>0.21</td>
<td>0.67 (0.06)</td>
<td>0.69 (0.10)</td>
</tr>
<tr>
<td>Threaten another resident with a cane, fist, or other object</td>
<td>0.81 (0.05)</td>
<td>0.74 (0.78)</td>
<td>0.25</td>
<td>0.25</td>
<td>0.70 (0.71)</td>
<td>0.42</td>
<td>0.70</td>
<td>0.58 (0.07)</td>
<td>0.02 (0.16)</td>
</tr>
<tr>
<td>Other physical behavior like kicking, biting, scratching, or spitting at another resident</td>
<td>0.80 (0.06)</td>
<td>0.63 (0.40)</td>
<td>0.77</td>
<td>1.00</td>
<td>0.72 (0.54)</td>
<td>0.66</td>
<td>0.96</td>
<td>0.61 (0.08)</td>
<td>0.74 (0.11)</td>
</tr>
<tr>
<td>Going into another resident's room without asking/taking/touching/damaging/breaking other residents' &quot;personal&quot; things</td>
<td>0.51 (0.06)</td>
<td>0.50 (0.71)</td>
<td>0.27</td>
<td>0.35</td>
<td>0.47 (0.74)</td>
<td>0.21</td>
<td>0.30</td>
<td>0.59 (0.08)</td>
<td>−0.23 (0.18)</td>
</tr>
<tr>
<td><strong>Eigenvalues</strong></td>
<td>5.2</td>
<td>1.2</td>
<td>1.4</td>
<td>0.6</td>
<td>4.9</td>
<td>1.8</td>
<td>1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correlation of scores with factors</strong></td>
<td>0.88</td>
<td>0.78</td>
<td>0.93</td>
<td>0.85</td>
<td>0.77</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Notes:* G = general factor; F1, F2, and F3 = group factors; $h^2$ = communality; (uniqueness $u^2 = 1$−communality).

$^a$Geomin (oblique) rotation.

$^b$S–L PC = S–L principal components.
between responses to a pair of items is greater than that predicted by the model, the values are flagged indicating that a cluster of items may measure an unmodeled dimension. Values that are larger than 10 are considered problematic (Thissen, 2011).

**Software.**—The ECV and McDonald’s omega were estimated using both exploratory and confirmatory factor analyses. R software (R Development Core Team, 2008) was used to calculate McDonald’s omega statistics, as well as several others recommended by Revelle and Zinbarg (2009), contained in the “psych” package that they developed in R (www.R-project.org). Polychoric and polyserial correlations based on the underlying continuous normal variables (Zi) were estimated using a SEM package, MPLUS (Muthén & Muthén, 1998–2010). The S–L solution was obtained in the R “psych” package and loadings were estimated using MPLUS software. Item Response Theory Patient Reported Outcomes (IRTPRO) software (Cai, Thissen, & du Toit, 2011; Thissen, 2011) was used for IRT parameter estimation and tests of model fit. IRTPRO incorporates various tests of LD, for example, SS-$\chi^2$ (Orlando & Thissen, 2003). The IBM Statistical Package for the Social Sciences (IBM SPSS, 2010) reliability module calculates the corrected item-total correlations and overall classical test theory estimate of reliability of the item set.

**Results**

The analytic sample of residents whose behavior was evaluated by the staff at baseline was 1,812. The majority (72.1%) were women; the mean age was 84.1 ($SD = 10.0$) years; the mean educational level was 11.7 ($SD = 4.1$) years; and 19.1% were Black, 16.6% were non-White Hispanic, 61.9% were non-Latino White, and 2.7% other.

**Preliminary Analyses**

Only the resident-to-resident mistreatment items were included in the psychometric analysis (Table 1). The frequency of specific incidents was low (Table 3). Therefore, the decision was made to combine any occurrence of a specific type of R-REM within the past 2 weeks or past year. Among the items in the final analysis, the highest frequency was 0.09 for the item “scream” and the next highest was 0.08 for the item “use bad words.” Five items with prevalence rates of 0.01 were excluded: “insulting race or ethnic group,” “saying sexual things,” “doing sexual things in front of another resident,” “touching in a sexual manner,” and “getting help when didn’t ask or want help.” Prior to combining items into testlets, there were some instances of local dependencies, for example, one pair of items showed very high LD statistics: “going to another resident’s room without asking” and “touch, damage, or break another resident’s personal things” (61.0). Thus, four items were combined in two testlets based on the LD results: “kicking,” “biting,” “scratching,” and “spitting” were combined and “going into room without asking,” “taking/touching things,” and “damaging things” were combined. The final analytic item set included 11 items (Table 1). After the items were combined, all item pair LD statistics were within the normal range.

**EFA Analyses**

The simple (4 to 1) eigenvalue rule provided some evidence for essential unidimensionality (eigenvalues of 6.943 to 1.264 or 5.5 to 1). The first component (eigenvalue) explained 63% of the variance (Supplementary Table 1). The scree plot showed a clear dominance of the first component (factor) relative to the other (Supplementary Figure 1). The fit indices for the unidimensional and two-factor models performed by MPLUS were CFI: one factor = 0.979 and two factor = 0.995. The RMSEAs were as follows: one factor = 0.030 and two factor = 0.017 (Supplementary Table 1). Because the eigenvalue ratio was moderate, further tests were performed using the bifactor model.

**Bifactor Model**

Prior to analyses, the inter-item tetrachoric correlation matrix was reviewed; correlations range from 0.314 to 0.878 (Supplementary Table 2). The graphic depiction of the bifactor model tested is shown in Supplementary Figure 2. The RMSEA statistic produced by “psych” R was 0.04 for the three-factor solution, 0.05 for the two-factor S–L solution, and 0.01 for the MPLUS bifactor model, applied using the general and two-group factor results from the S–L solution (Table 2). The MPLUS bifactor models with three-group factors did not converge. Table 1 shows the factor loadings from the S–L solution and from the MPLUS solutions for both the unidimensional and bifactor models. The relative size of the general and group factor loadings (MPLUS two-group factors) indicates a very strong
items loaded on a third S–L group factor: “threaten another resident with cane, fist, or other object”; “throw things at another resident”; and the testlet item “going to another resident’s room without asking or taking, touching, damaging, or breaking another resident’s personal things.” However, as stated earlier, the MPLUS three-group factor solution did not converge; thus, there is robust support for at most two-group factors.

The IRT discrimination (a) parameters estimated with IRTPRO, ranged from the lowest 1.18 (going to another resident’s room without asking or take, touch, damage, or break another resident’s personal things) to the highest 4.95 (scream; Table 3). Additional items with high discrimination parameters greater than 3.0 were as follows: “use bad words” (4.16); “threaten another resident with cane, fist, or other objects” (3.78); “try to scare, frighten, or threaten with words” (3.63); and “throw things at another resident” (3.52). The severity (difficulty) parameters (b) ranged from 1.43 (scream), the least severe item, to 3.30 (going to another resident’s room without asking/take, touch, damage, or break another resident’s personal things; Table 3). As expected, the verbal items were less severe indicators of R-REM.

The information functions (estimated using IRTPRO) for all the items are located at the upper half of the theta continuum; there was little information contributed at the lower half of the underlying construct distribution (Figure 1). The highest information function (5.2) was for the item “scream at another resident,” which peaked at the theta level of 1.6 followed by the item “use bad words” (4.3) at theta 1.6. The items “threaten another resident with cane, fist, or other objects”; “try to scare, frighten, or threaten with words”; and “throw things at another resident” were more informative at even higher theta
levels (the highest information points: 3.2 at theta level 2.4, and 3.2, 2.7 all at theta level 2.8). The testlet item “going to another resident’s room without asking or take, touch, damage, or break another resident’s personal things” provides minimum information (highest information point: 0.3 at theta level 2.8; Figure 1 and Supplementary Figure 3).

The essential dimensionality estimates were high; the ECV from the “psych” R program was 0.64 and 0.76 from the MPLUS bifactor model. Different reliability estimates were obtained from several methods. The alpha estimate from the “psych” R package was 0.94, omega hierarchical (ω_h; three-group solution) was 0.76, and omega total (ω_t; two-group solution) was 0.97 (Table 2); thus, adding support for a two- rather than three-group solution. The Cronbach’s alpha reliability estimate was 0.74 and standardized alpha was 0.75. Corrected item-total correlations ranged from 0.30 (grab or yank another resident) to 0.61 (scream; Supplementary Table 3). Based on the IRT-PRO results, the estimates of reliability were calculated along the theta continuum and ranged from 0.50 (lower end of theta) to 0.95 at theta 2.4 (Supplementary Table 4). Reliabilities were adequate (≥ 0.90) at θ ≥ 1.2.

Verbal and physical subscales were examined using classical test theory reliability estimates. The combined verbal items (use bad words toward another resident, scream at another resident, try to scare another resident with words, and boss around, tell people what to do) resulted in a Cronbach’s alpha estimate of 0.73 and 0.74 for unstandardized and standardized versions, respectively, with corrected item-total correlations ranging from 0.48 to 0.66. The six physical items (hit another resident; grab another resident; push another resident; throw things at another resident; threaten another resident with cane, fist, or other object; and other physical behavior like kicking, biting, scratching, or spitting at another resident) resulted in a Cronbach’s alpha estimate of 0.65 (for both unstandardized and standardized alphas). The corrected item-total correlations ranged from 0.31 to 0.48. It is noted that the factor analyses placed the item, “threaten another resident with cane, fist, or other object” in the verbal as well as the third factor along with the items “throw things at another resident” and “going to another resident’s room without asking”/“take, touch, damage, or break another resident’s personal things” in the three-group S–L solution, even though the content was verbal threatening of physical actions. Placing the threatening item in the verbal scale reduces slightly the internal consistency estimate of the scale (from 0.73 to 0.71) but increases the alpha estimate if the item is added to the physical subscale (from 0.63 to 0.65). The internal consistency estimate did not change whether the item, “throw things at another resident” was included in the physical subscale (0.63). Thus, one could plausibly use a four-, five-, or six-item subscale for each, if desired. However, the final confirmatory bifactor results shown in Table 1 suggested the possibility of four item measures of physical and verbal mistreatment. It is emphasized that the bottom-line result is that the factor analyses supported most strongly an essentially unidimensional measure.

### Table 3. Item Response Theory (IRT) Item Parameters and Standard Error Estimates (IRTPRO)

<table>
<thead>
<tr>
<th>Item description</th>
<th>Base rate (% positive)</th>
<th>a</th>
<th>SE of a</th>
<th>b_1</th>
<th>SE of b_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use bad words toward another resident</td>
<td>7.7</td>
<td>4.16</td>
<td>0.50</td>
<td>1.57</td>
<td>0.06</td>
</tr>
<tr>
<td>Scream at another resident</td>
<td>9.3</td>
<td>4.95</td>
<td>0.68</td>
<td>1.43</td>
<td>0.05</td>
</tr>
<tr>
<td>Try to scare, frighten, or threaten (another resident) with words</td>
<td>1.9</td>
<td>3.63</td>
<td>0.58</td>
<td>2.29</td>
<td>0.12</td>
</tr>
<tr>
<td>Boss around/tell another resident what to do</td>
<td>4.5</td>
<td>2.63</td>
<td>0.30</td>
<td>2.06</td>
<td>0.11</td>
</tr>
<tr>
<td>Hit another resident</td>
<td>2.0</td>
<td>2.60</td>
<td>0.38</td>
<td>2.48</td>
<td>0.16</td>
</tr>
<tr>
<td>Grab or yank another resident</td>
<td>1.3</td>
<td>2.24</td>
<td>0.39</td>
<td>2.88</td>
<td>0.26</td>
</tr>
<tr>
<td>Push or shove another resident</td>
<td>1.5</td>
<td>2.32</td>
<td>0.38</td>
<td>2.76</td>
<td>0.22</td>
</tr>
<tr>
<td>Throw things at another resident</td>
<td>1.0</td>
<td>3.52</td>
<td>0.69</td>
<td>2.58</td>
<td>0.17</td>
</tr>
<tr>
<td>Threaten another resident with a cane, fist, or other object</td>
<td>0.8</td>
<td>3.78</td>
<td>0.84</td>
<td>2.63</td>
<td>0.19</td>
</tr>
<tr>
<td>Other physical behavior like kicking, biting, scratching, or spitting at another resident</td>
<td>1.0</td>
<td>2.34</td>
<td>0.44</td>
<td>2.92</td>
<td>0.27</td>
</tr>
<tr>
<td>Going into another resident’s room without asking or taking/touching/damaging or breaking other resident’s “personal” things</td>
<td>3.6</td>
<td>1.18</td>
<td>0.19</td>
<td>3.30</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Examining the distributional characteristics of the measure, the mean sum score was 0.35 ($SD = 0.99$) with the median of 0.0; the mean theta was 0.002 (0.62) with the median of −0.26. Skewness statistics were 3.84 (0.06) and 2.27 (0.06) and kurtosis was 17.95 (0.12) and 3.88 (0.12). The Kolmogorow–Smirnov test of normality provided evidence of highly skewed distributions (20.36, $p < .001$ and 21.38, $p < .001$; Supplementary Table 5). The majority of the cases had the sum score of zero (84.3%), which is equivalent to the theta value −0.26. Examining the expected item scores mapped onto theta, the highest score for the sample was 9, which is equivalent to a theta value of 2.92. The highest theoretical value for the sum score is 11 corresponding to an estimated theta value of 3.4; however, no cases had the sum score 10 or 11. There is almost a complete overlap of the cumulative proportion for the sum score and the modeled proportion (Table 4).

**Discussion**

The study supports the unidimensional assessment of R-REM using the 11-item scale; however, modest support was provided for the use of separate subscales for verbal and physical constructs. The S–L bifactor model identified three factors: verbal, physical, and a less differentiated factor including items on room invasion and threatening gestures; however, this result was not supported by the final estimation procedures.

**Limitations**

A limitation of these analyses is the low prevalence of many items, resulting in omission of items that may be potentially salient and stressful, for example, sexual encounters. Thus, insertion of a combined item is recommended: “saying sexual things and inappropriate touching of another resident.” The proposed 12-item measure is appended. Additionally, items were combined into presence in the past 2 weeks or year. Although this method may be best for analyses of data and for constructing global summary scales, for individual assessments, it may be desired to collect data separately on recent and past R-REM events to obtain a frequency of occurrences.
measure to inform appropriate care planning and individual interventions. Thus, the items for up to five incidents occurring within the past 2 weeks and two additional incidents occurring within the past 1 year are included in the measure shown in Supplementary Material. Finally, the distributions of R-REM were skewed and higher reliability estimates were observed at the more severe tail of the R-REM distribution. Reliability estimates are lower at the less severe tail of the distribution where noncases are prevalent. In part, this is due to the low base rate of R-REM. It is not surprising that greater accuracy is observed at the “caseness” tail of the distribution, where reporting data are available in contrast to the “noncase” tail, where the potential for error is most likely due to failure to observe or report R-REM. Nonetheless, it is important to have high reliability among cases because these are the ones singled out for adjudication and appropriate actions. Moreover, the phenomenon of higher reliability in the middle and upper section of the theta distribution is not uncommon in the field of patient-reported outcome data. For example, this phenomenon is commonly observed in data from the National Institutes of Health Patient Reported Outcome Measurement Information System (PROMIS; Reeve et al., 2007; Teresi et al., 2009).

**Conclusion**

Most studies of aggressive behavior have used staff informant reports or observations. Creative methodologies have typically been developed and employed in studies of behavioral disturbance in nursing home residents. Some of the observational methods that have been used for this purpose include direct videography of groups of residents for brief proscribed periods (Kihlgren et al., 1993), bar-coding strategies in which researchers tally behaviors of interest with a light pen system (Bridges-Parlet, Knopman, & Thompson, 1994; Holmes & Teresi, 1996), and computer-assisted observations in which researchers use a laptop computer to keystroke appropriate tabs when residents are observed in specific activities (e.g., during morning care; Rogers, Holm, & Burgio, 1999). It was determined that none of these methods were practical or feasible for four reasons. First, episodes of R-REM are likely to be too brief and too intermittent to make any of these methods useful. A large research staff would have to be positioned virtually everywhere in the facility for extensive periods to make meaningful observations; this would be costly and potentially invasive for staff and residents, and most importantly, not sustainable beyond the life of a research project. Second, episodes of R-REM are unpredictable. In arriving at this methodology, the investigators also considered oversampling suspected high-risk times and environments (e.g., congregate meals), but no data exist on the epidemiology of when R-REM occurs on which to base such a sampling strategy. Third, violent episodes that are volitional are probably less likely to occur with research observers monitoring residents directly; their presence could actually influence the prevalence of R-REM. Finally, the use of videography

<table>
<thead>
<tr>
<th>Summed score</th>
<th>N</th>
<th>%</th>
<th>Cumulative %</th>
<th>EAP(^a) [θ/x]</th>
<th>SD [θ/x]</th>
<th>Modeled %</th>
<th>Modeled Cumulative %</th>
<th>Reliability(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1,526</td>
<td>84.3</td>
<td>84.3</td>
<td>-0.255</td>
<td>0.834</td>
<td>83.9</td>
<td>83.9</td>
<td>0.52</td>
</tr>
<tr>
<td>1</td>
<td>109</td>
<td>6.0</td>
<td>90.3</td>
<td>0.902</td>
<td>0.569</td>
<td>8.35</td>
<td>22.2</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>4.8</td>
<td>95.1</td>
<td>1.466</td>
<td>0.326</td>
<td>3.27</td>
<td>95.5</td>
<td>0.92</td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>2.4</td>
<td>97.5</td>
<td>1.736</td>
<td>0.283</td>
<td>1.86</td>
<td>97.4</td>
<td>0.93</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>1.1</td>
<td>98.6</td>
<td>1.952</td>
<td>0.273</td>
<td>1.06</td>
<td>98.4</td>
<td>0.93</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>0.7</td>
<td>99.3</td>
<td>2.153</td>
<td>0.267</td>
<td>0.61</td>
<td>99.0</td>
<td>0.93</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>0.5</td>
<td>99.8</td>
<td>2.345</td>
<td>0.262</td>
<td>0.37</td>
<td>99.4</td>
<td>0.94</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.1</td>
<td>99.9</td>
<td>2.531</td>
<td>0.263</td>
<td>0.24</td>
<td>99.7</td>
<td>0.95</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.1</td>
<td>100</td>
<td>2.719</td>
<td>0.274</td>
<td>0.16</td>
<td>99.8</td>
<td>0.94</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.1</td>
<td>100</td>
<td>2.923</td>
<td>0.299</td>
<td>0.10</td>
<td>99.9</td>
<td>0.94</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0.0</td>
<td>100</td>
<td>3.159</td>
<td>0.345</td>
<td>0.06</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0.0</td>
<td>100</td>
<td>3.426</td>
<td>0.402</td>
<td>0.02</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Notes: \(^a\)EAP is the expected a posteriori estimate (mean of the posterior distribution of theta \([\theta]\), conditional on \(x\) [the observed response pattern]).

\(^b\)The average reliability across all measured levels of theta (\(\theta\)) is 0.67.
was ruled out as too invasive for sustained use, particularly in resident’s rooms. Although the experience with bar-code technology has been positive, it is not likely that this method can be implemented on a wide scale basis without extensive support. However, the increased use of electronic medical records could make this approach viable in the future.

Given the previous considerations, this project adopted a strategy of R-REM case identification that makes use of the best potential reporters—residents themselves and the staff who care for them. Analogously, the instrument proposed to measure the phenomenon derives its methodology from two fields (and representative scales reflective of them) that have proven strategies appropriate for this study—the CTS (derived from the interpersonal aggression literature) and the CMAI (derived from the nursing home behavioral disturbance literature).

A challenge in long-term care settings is recognition and documentation of resident-to-resident mistreatment. Because of the ubiquity of aggressive behaviors in general as well as towards residents (Castle, 2012) and staff (Lachs et al., 2012; Morgan et al., 2012), R-REM may often be ignored. This may be due in part to desensitization because such incidents may be perceived as daily routine events. Although staff training has resulted in increased recognition and reporting (Teresi et al., 2012), documentation is still a challenge. Documentation of aggressive behaviors toward residents is not only important in and of itself for record keeping, but it can also be crucial as a preventive measure at an individual level for risk management and care planning. Additionally, such documentation is important at the institutional level to avoid financial and or licensing penalties (Soreff, 2012). Thus, methods for increasing recognition, reporting, and documentation benefit residents, direct care, and administrative staff.

Castle (2012) has called for measurement approaches at the individual resident level that can provide representative estimates of incidence and prevalence of R-REM. Additionally, such a measure could be incorporated into routine data collection used for care planning. Collectively, it is hoped that the extensive qualitative and quantitative work that resulted in the development of this measure will help to advance the measurement and ultimately interventions associated with this important and underrecognized problem facing residents of long-term care settings.

**Supplementary Material**

Supplementary material can be found at: http://gerontologist.oxfordjournals.org.

**Funding**

This work was supported in part by National Institute on Aging (AG014299-06A2), National Institute of Justice (FFO 42USC3721), New York State Department of Health Dementia Grant Program (contract # C-022657). This work does not necessarily reflect the opinions of the funding organizations.

**References**


Teresi, J. A., Ramirez, M., Ellis, J., Silver, S., Boratgis, G., Kong, J., ... Lachs, M. S. (2012). A staff intervention targeting resident-to-resident elder mistreatment (R-REM) in long-term care increased staff knowledge, recognition and reporting: Results from a cluster randomized trial. *International Journal of Nursing Studies*.
