Rasch Analysis of Lebanese Nurses' Responses to the Occupational Fatigue Exhaustion Recovery Scale

Michael Clinton, PhD, RN

Hariri School of Nursing, American University of Beirut, Lebanon

Hasmig Tchaparian, MSN, RN

American University of Beirut Medical Center, Lebanon

Background and Purpose: The purpose of our methodological study was to explore the psychometric properties of Occupational Fatigue Exhaustion Recovery (OFER) subscales. Methods: Rasch analyses of 366 Lebanese registered nurses' responses to the Chronic Fatigue (CF), Acute Fatigue (AF), and Intershift Recovery (IR) subscales. Results: Disordered rating categories, response dependence, and possible differential item functioning (DIF). The data were a better fit to a 3-dimensional Rasch rating scale model; difference, $\chi^2 = 104$, $df = 12$, $p = .01$; unidimensional Akaike information criterion (AIC) = 11,925; multidimensional AIC = 11,821. Conclusion: Multidimensional analysis confirmed that the CF and AF subscales have sufficient reliability for use in exploratory studies of fatigue with English-speaking respondents in the Eastern Mediterranean region. An Arabic version of OFER subscales is required to facilitate future studies in Lebanon and the region. Norm values are reported to facilitate international comparisons.

Keywords: psychometrics; registered nurses; fatigue; Lebanon

As health care systems respond to increasing levels of acuity, the impact of chronic illnesses, and the burden of disease on national fiscal policies, occupational fatigue among health care workers has become an imperative for investigators. Whereas long shifts, insufficient time for recovery between shifts, and pressures of work–life balance are common causes of nurse fatigue internationally, country-specific aspects require investigation if local nursing shortages are to be overcome.

In Lebanon, a small country on the eastern shore of the Mediterranean Sea with a population of 4 million people and more than 2 million Syrian and Palestinian refugees, organizational factors intensify pressures on the nursing workforce. Public hospitals have a nursing shortage, equipment is often in short supply, and salaries are sometimes paid late. Dispensaries and charities provide services in areas without hospitals. Academic medical centers can recruit well-qualified nurses, but retention is difficult because the Gulf States offer generous employment packages and better working conditions. The nursing shortage that results is made worse by talented nurses leaving Lebanon to pursue careers or PhD studies in the United States, Canada, or Europe. Health care organizations respond by introducing 12-hr shifts, compulsory overtime, and calling in off-duty nurses. Longer work

http://dx.doi.org/10.1891/1061-3749.25.3.459

periods between days off and requiring nurses to alternate between day and night shifts at short notice are other staffing practices. Although contrary to policy, some medical center nurses have a second nursing job.

Investigation of the relationship between the organization of nursing work and occupational fatigue has become a research imperative in Lebanon. Studies are required modeled on those conducted in other countries (Barker & Nussbaum, 2011; Brooks, 2000; Chana, Kennedy, & Chessell, 2015; Chau, West, & Mapedzahama, 2014; Chen, Davis, Daraiseh, Pan, & Davis, 2014; Eriksen, 2006). Furthermore, the long-term health consequences of work strain fatigue among nurses requires investigation (Choi et al., 2014; Heikkilä et al., 2013; Heikkilä, 2012; Kivimäki et al., 2012). Another priority is research that investigates the relationship between nurse fatigue and patient safety (Chana et al., 2015; Kunaviktikul et al., 2015; Samra & Smith, 2015). Consequently, valid and reliable instruments are essential to measure fatigue and the relationships between occupational fatigue, the health status of nurses, and patient safety. Equally important is research that examines the impact of national health policies, funding models, organizational cultures, and leadership behaviors on occupational fatigue among nurses.

The Occupational Fatigue Exhaustion Recovery (OFER) scale (Rella, Winwood, & Lushington, 2009; Winwood, Lushington, & Winefield, 2006; Winwood, Winefield, & Lushington, 2006) has been used recently in the first national study of occupational fatigue among nurses in Lebanon. We conducted a preliminary study with a convenience sample of 366 registered nurses fluent in English at an academic medical center in Beirut. We used the data from this study, consecutive Rasch analyses, and an exploratory multidimensional analysis to examine the psychometric characteristics of the OFER subscales. We will report the findings of the national survey separately.

THE RASCH MEASUREMENT MODEL

The Rasch measurement model was developed by the Danish mathematician Georg Rasch (1960), modified and applied by Ben Wright (Wright, 1977; Wright & Stone, 1979), and later extended for polytomous data by David Andrich (1978). Wright and Masters (1982) extended Rasch analysis to attitude questionnaires and described estimation procedures for ordered category data. Masters (1982) introduced the partial credit model as an extension of the Andrich rating scale model. The partial credit model is used when the number of rating categories varies for items in the same rating scale. Rasch analysis is a special case of the general linear model.

The Rasch model estimates the probability of endorsing an item in each rating scale category. The probability of endorsement and the propensity to endorse an item is measured in logits (additive log-odds units of equal measurement) on the same continuous latent variable. The expected probability of endorsing one of the two categories on a dichotomous rating scale is 0.5. Which category is endorsed depends on the location of the respondent on the trait. The respondent's trait measure similarly governs the probability of endorsing an item in each of the ordered categories in a polytomous rating scale. Data for polytomous rating scales are fit to the following mathematical model: $\text{loge}(P_{nij} / P_{ni}(_{j}\text{-}1)) = B_n - D_i - F_j$.

Loge is the natural logarithm of the probability P_{nij} of respondent $_n$ of ability B_n endorsing category $_j$ in response to a scale item of difficulty D_i, as opposed to the probability $P_{ni}(_{j}\text{-}1)$ of the respondent endorsing the item in the next lowest category $(_{j}\text{-}1)$. The parameter F_j is the

Rasch–Andrich threshold, the point on the latent variable corresponding to where the probability curves for adjacent rating categories cross. This is the point where there is equal probability of endorsement in either of the two adjacent categories. Dichotomous items have only one Rasch–Andrich threshold. Polytomous items have $_k$-1 Rasch–Andrich thresholds where $_k$ is the number of rating scale categories.

The Rasch model is one of four logistic models commonly fit to rating scale data. The number of parameters estimated varies from one to four. The *b* parameter is the item location parameter or point of inflection on the latent variable scale. The *a* parameter is the item discrimination parameter; the slope of the item characteristic curve (ICC) corresponding to the point of inflection. The *c* parameter is the lower asymptote used to estimate the selection of correct answers by guessing in tests of ability. The *u*, or "carelessness" parameter, is the upper asymptote of the ICC.

The Rasch model estimates the *b* parameter only; the two-parameter model estimates *a* and *b*; the three-parameter model, *a*, *b*, and *c*; and the four-parameter model, all four parameters. The models share the assumption that one underlying latent variable measures the trait under investigation. The latent trait is measured in logits from $-\infty$ to ∞. Latent measures in the range -3–$+3$ logits centered on 0 logits are sufficient for most purposes.

The Rasch measurement model is the model of choice when investigating the psychometric characteristics of rating scales because it examines the spread of item and respondent locations on the same latent variable, calibrates measurement error to improve precision, indicates the probability of item fit, and requires that the data fit the model and not that the model fit the data (Granger, 2008). The property of *conjoint additivity* that enables respondents and items to be located on the same linear latent variable is unique to Rasch measurement.

The raw scores of respondents to a rating scale are *sufficient statistics* to locate them and rating scale step thresholds on the latent trait. The underlying assumption is that respondents with more of the latent trait will endorse items in higher rating categories. Step thresholds are the locations on the latent variable at which the probability of endorsing an item in one of two adjacent categories is equal and do not vary across items. However, the distance between step thresholds is not assumed to be equal for a particular item. The assumption that raw scores are sufficient statistics to estimate Rasch parameters is a consequence of the property of *constant item discrimination*, which constrains the *a* parameter to equality and rules out the possibility that respondent trait measures can vary from item to item.

Separability enables item parameters to be estimated without knowing the distribution of the latent trait among respondents. Separability gives Rasch measurement its specific *objectivity*, the property of complete independence of item and respondent measures. Specific objectivity ensures that respondent measures are independent of the items used if the items are well defined and homogeneous in measuring the latent trait (Rost, 2001). Similarly, subject to a large heterogeneous sample of respondents, the results of Rasch analyses can be extrapolated to the population of interest (Granger, 2008).

The property of *latent additivity* is central to Rasch analysis because it requires that addition or subtraction is used to connect respondent and item measures on the log-linear latent variable. Latent additivity enables comparison of the difficulty of endorsing items in successively higher categories and the measured traits of respondents.

Rasch properties maximize the homogeneity of the latent trait by allowing redundant items to be removed without sacrificing measurement information (Granger, 2008). However, both the strength and weakness of the Rasch measurement model is the assumption that the interaction between respondents and items in the variation of latent measures is given only by the difficulty of the items, the traits of the respondents measured by the item set, and the rating scale structure of the research instrument (Granger, 2008).

Although the simplest of the four logistic models, the Rasch measurement model has been widely adopted and has considerable application in the human sciences (Boone, Staver, & Yale, 2014). Our choice of the Rasch measurement model as the starting point for our analyses is further justified because items with Likert or other polychromous rating scales require large samples to achieve stable estimates of *a*, *c*, and *u*.

RASCH FIT STATISTICS

Rasch fit statistics indicate how closely respondents and their responses match the pattern predicted by the Rasch measurement model. Inlier-pattern sensitive (infit) and outlier-sensitive fit (outfit) statistics are calculated as mean squared (MNSQ) values. Infit and outfit values for an item that perfectly matches the Rasch measurement model have an MNSQ of 1. Items with MNSQ values >1 overfit the model and lack precision. MNSQ values <1 indicate that responses to an item are too predictable and may not contribute to successful measurement. Outfit and infit MNSQs in the range 0.77–1.3 are acceptable for most purposes (Linacre, 2015a). An alternative rule of thumb is to accept MNSQ values in the range 0.6–1.4 (Frantom, Green, & Hoffman, 2002). For exploratory analysis, a range of 0.5–1.5 is acceptable (Linacre, 2002). Adjusted for sample size (Smith, Schumacker, & Bush, 1998), the acceptable range of MNSQ values for this study was 0.68–1.32.

METHODS

Aim of the Study

Our aim was to explore the psychometric properties of the OFER subscales. To achieve our aim, we undertook secondary analyses of data collected for a cross-sectional survey designed to investigate the relationship between professional quality of life and fatigue among nurses at an academic medical center in Beirut. We examined the unidimensionality and Rasch fit statistics for the OFER Chronic Fatigue (CF), Acute Fatigue (AF), and Intershift Recovery (IR) subscales and the possibility of item bias. We used a multidimensional rating scale model with three dimensions to further investigate the dimensionality of the OFER subscales.

Sample

All registered nurses involved in direct patient care at an academic medical center in Beirut were eligible to participate in the survey. The institutional review board, the medical director, and the director of nursing approved the study. We posted fliers at nurses' stations and visited clinical units and departments to explain the study and leave survey packages. Respondents confirmed voluntary informed consent by returning completed questionnaires to conveniently located drop boxes. Our convenience sample of 366 respondents was recruited from a nursing workforce of 450 nurses (response rate 81%). We describe the characteristics of the sample in Table 1.

TABLE 1. Sample Characteristics

	n	%
Age (years)		
≤24	57	15.5
25–35	253	68.8
36–45	42	11.4
46–50	14	3.8
Gender		
Female	231	62.8
Male	134	36.9
Marital status		
Single	206	56.0
Married	146	39.7
Other	14	3.8
Education		
BS in nursing	259	70.4
Master's in nursing	83	22.7
Other	24	6.6
Enough rest		
Yes	152	41.3
No	210	57.1
Obliged to work		
Yes	128	34.8
No	237	64.4
Second job		
Yes	45	12.2
No	320	87.4

Note. BS = bachelor of science.

Instrument

The 15-item OFER scale was developed to measure work-related fatigue (Winwood, Lushington, et al., 2006). A 7-point rating scale (0–6) is used to ensure sufficient sensitivity. Subscale scores for CF, AF, and IR are converted to quotients. Winwood, Lushington, et al. (2006) noted that negatively keyed items can result in errors caused by respondent carelessness and that scales that have only positively keyed items lead to artificial factor solutions and may lack unidimensionality. The OFER has 5 negatively keyed items and 10 positively keyed items. The OFER is unique in measuring IR. Winwood, Lushington, et al. (2006) report that the gender neutrality of the OFER was confirmed in a pilot study conducted on female nurses and male quarry workers.

DATA ANALYSIS

We used the Rasch rating scale model to conduct consecutive analyses of the measurement performance of the three OFER subscales. All unidimensional analyses were conducted with WINSTEPS version 3.92.0 (Linacre, 2015b). We examined response ordering for the three subscales. Principal components analysis (PCA) of residuals was used to identify possible non-Rasch dimensions. Inter-item standardized correlations of ≥.30 were taken as evidence of item dependence. Outfit and infit MNSQ were examined. Outfit MNSQ indicates the discrepancy between an observed and a Rasch expected response irrespective of the distance between the response and respondent measure on the latent trait. Infit MNSQ indicates an unexpected response near to the respondent's latent trait measure (Linacre, 2015a). We regarded an item as too imprecise if sample adjusted outfit MNSQ values were >1.32 and as overly predictable if they were <0.68. We examined respondent separation indices and item reliability coefficients to assess subscale precision. We looked for respondent separation indices ≥2.0 and item reliability coefficients of ≥.80 (Linacre, 2015a). Finally, we examined pairwise differences between groups to identify differential item functioning (DIF).

RESULTS

Chronic Fatigue Subscale—Five Items

Global Statistics. The Rasch rating scale model was an acceptable fit to the data: log likelihood chi-square = 4,175.65, *df* = 4,220, *p* = .65, and root mean square standard error (RMSE) = 0.9007.

Misfitting Respondents. We identified 32 respondents with outfit MNSQ values >2.0 but included their data in our analyses to ensure adequate representation of the sample.

Response Ordering. Rating scale Categories 1 (*disagree*) and 2 (*slightly disagree*) and Rasch–Andrich thresholds for the intersections between Categories 2 and 3 (*slightly disagree* and *neutral*) and Categories 3 and 4 (*neutral* and *slightly agree*) were disordered. We did not correct category disordering.

Response Dependence. We examined largest standardized residual correlations to identify dependent items in the CF subscale and found none ≥.03.

Item Location. Respondents' mean CF raw scores were in the range 4.05 (*n* = 366) for Item 2 to 4.62 (*n* = 336) for Item 5. When measured on the Rasch dimensioned propensity to express CF, mean respondent measures ranged from −0.40 logits for Item 5 to 0.31 logits for Item 2.

Respondent Measures and Item Locations. Because the mean respondent measure for the CF items was 1.41 logits with the mean location of items on the latent trait scaled to 0.0, it was easy for this sample of respondents to endorse CF items in higher categories.

Item Fit Analyses. All five CF items were in our sample adjusted range for MNSQ values (0.68–1.32).

Dimensionality. The Rasch measures explained 59.5% of the raw score variance in the data (Table 2). The total raw unexplained variance for this sample was 40.6% (first five contrasts). The eigenvalue for the first contrast (1.8) was <2.00, considered acceptable for the proportion of unexplained variance predicted by the Rasch model.

TABLE 2. Rasch Analysis Fit Statistics for Occupational Fatigue Exhaustion Recovery Subscales

	Items Examined No.	Variance Explained Rasch (%)	Person Separation Ratio	Mean Person Measure (logits)	Model *SE* (logits)	RMSE	Person Reliability	Principal Component Analysis (Eigenvalues)
Chronic Fatigue	5	59.5	2.18	1.41	0.62	0.71	.83	1.8
Acute Fatigue	5	54.6	1.29	0.56	0.39	0.49	.63	3.3
Positively scored items	3	73.2	2.27	5.38	1.28	1.51	.84	1.3
Intershift Recovery	5	35.3	1.29	0.20	0.34	0.36	.57	2.8
Reverse scored items	3	57.0	1.76	1.02	0.69	0.76	.76	1.8

Note. All estimates for nonextreme respondents. Real (inflated for misfit) values reported. Extreme person scores excluded. Variance explained—Rasch (%) = proportion of variance explained by Rasch measures; Model *SE* = model standard error; RMSE = root mean square average of the standard errors; Eigenvalues = eigenvalues for first contrasts.

Targeting. The CF items were a good match to the sample because the average respondent measure was less than one error of measurement from the average item measures scaled to 0.0 (Fisher, 2007). However, the CF subscale had considerable ceiling and floor effects. Whereas respondent measures ranged from −3.86 to 5.81 logits, item measures were in the narrow range −0.40 to 0.31 logits. Nevertheless, the CF items spread the sample and had an acceptable operational range of approximately 6.5 logits.

Respondent Separation Index and Respondent Reliability. The model respondent separation index of 2.18 indicated that the CF subscale had sufficient sensitivity to separate the sample into four strata of CF (*none*, *low*, *moderate*, and *high*). The CF subscale has acceptable respondent reliability (see Table 2). The approximate Cronbach's α for respondent raw score test reliability was .84.

Differential Item Functioning. We found no evidence of DIF.

Acute Fatigue Subscale—Five Items

Global Statistics. The Rasch rating scale model was a plausible fit to the data: log-likelihood chi-square = 5,279.83, *df* = 5,280, *p* = .50, and RMSE = 1.2364.

Response Ordering. Rating scale Categories 1 (*disagree*) and 2 (*slightly disagree*), Categories 2 and 3 (*slightly disagree* and *neutral*), and Categories 3 and 4 (*neutral* and *slightly agree*) and Rasch–Andrich thresholds for the intersections between Categories 2 and 3 (*slightly disagree* and *neutral*), Categories 3 and 4 (*neutral* and *slightly agree*), and Categories 4 and 5 (*slightly agree* and *agree*) were disordered. We did not correct category disordering.

Response Dependence. We identified standardized residual correlations of .53, .52, .41, and .37 and, therefore, response dependence for Items 6 and 7, Items 7 and 8, Items 9r and 10r, and Items 6 and 8.

Item Location. Respondents' mean AF raw scores were in the range 2.38 (*n* = 366) for Item 10r to 4.92 (*n* = 366) for Item 7. When measured on the Rasch dimensioned propensity to express AF, mean difficulty measures ranged from −0.59 logits for Item 7 to 0.86 logits for Item 10r.

Respondent Measures and Item Locations. Given that the mean respondent measure for this sample was 0.56 logits with the mean location of items on the latent trait scaled to 0.0, it was easy for the respondents to endorse the AF items in higher rating categories.

Item Fit Analyses. Item 10r (outfit MNSQ 1.43) overfit our item fit sample adjusted range (Smith et al., 1998) of 0.68–1.32.

Dimensionality. The Rasch measures explained 54.6% of AF raw score variance (see Table 2). There was evidence that the AF subscale is not unidimensional. The eigenvalue of 3.3 for the first contrast suggests that the three positively keyed items (6, 7, and 8) form a second dimension. This second dimension had a disattenuated correlation of −1.0 with the negatively keyed items. A negative disattenuated correlation implies that the relationship between the Rasch dimension and the second dimension is undefined but negative according to the data.

Targeting. The mean difficulty of the items was a good fit to the mean trait levels of the respondents. However, there were marked ceiling and floor effects. Respondent measures were in the range of −2.93 to 4.27 logits, whereas item measures ranged from −0.59 to 0.58 logits. The AF items were less effective in spreading the sample than the CF items.

Respondent Separation Index and Respondent Reliability. The AF subscale had insufficient sensitivity to separate the respondents into three strata and had moderate reliability (see Table 2).

Differential Item Functioning. We found slight to moderate evidence of DIF for Item 7 (effect size = 0.51 logits, $t = -3.60$, $df = 185$, $p = .0004$) in favor of respondents who were obliged to work.

Positively Keyed Items. The three positively keyed items explained a higher proportion of variance and had a more acceptable respondent separation index than the full AF subscale. However, the proportion of variance explained by the second dimension (15.3%) explained more variance in the data than the positively keyed items (12.8%). The fit of the items to the sample was poor (see Table 2).

Intershift Recovery Subscale—Five Items

Global Statistics. The Rasch rating scale model was a plausible fit to the data: log-likelihood chi-square = 5,998.56, $df = 5{,}998$, $p = .49$, and RMSE = 1.3734.

Response Ordering. All rating scale categories were disordered except for Category 6 (*strongly agree*). No Rasch–Andrich thresholds were disordered. We did not correct category disordering.

Response Dependence. We examined largest standardized residual correlations to identify dependent items and found possible dependency for Items 12 and 14 (standardized residual correlation .48).

Item Location. Respondents' mean IR raw scores were in the range 2.54 ($n = 366$) for Item 14 to 3.98 ($n = 366$) for Item 15. When measured on the Rasch dimensioned propensity to express IR, the mean difficulty measures ranged from −0.34 logits for Item 15 to 0.42 logits for Item 14.

Respondent Measures and Item Locations. The IR items were easy for the respondents to endorse (mean trait measure 0.20, mean item location scaled to 0.0).

Item Fit Analyses. None of the items had an outfit MNSQ value outside our sample adjusted range (Smith et al., 1998) of 0.68–1.32.

Dimensionality. The Rasch measures explained 35.3% of the IR raw score variance (see Table 2). The proportion of variance explained by the first contrast (36.4%) was greater than that explained by the Rasch measures and more than that explained by the items (28.1%). The two positively keyed items formed the separate cluster. The disattenuated correlation between the positively and negatively keyed items was −.76. Therefore, the relationship between the Rasch dimension and the second dimension is undefined but negative according to the data.

Targeting. The IR items were an excellent fit to the Rasch rating scale. However, ceiling and floor effects were evident. Respondent measures were in the range −1.22 to 4.88 logits, whereas item measures were in the range −0.34 to 0.42 logits. Consequently, the IR items were only precise when measuring respondents with moderate levels of IR.

Respondent Separation Index and Respondent Reliability. The model respondent separation index of 1.16 indicated that the IR subscale had poor respondent reliability (.57) and insufficient sensitivity to separate this sample into more than two strata (see Table 2).

Differential Item Functioning. We found no evidence of DIF for IR items.

Positively and Negatively Keyed Items. We examined the Rasch measurement characteristics of the negatively keyed items. The proportion of variance explained; the respondent separation index and respondent reliability increased. The three items were not well matched to the sample (see Table 2).

CATEGORY REDUCTION

We tested the proposition that any improvement in data fit to the Rasch model would be offset by lower reliability by examining the effect of category reduction in consecutive unidimensional analyses of the OFER subscales. As expected, there was a decrease in model person reliability for all three subscales: from .83 to .82 for the CF subscale, from .63 to .51 for the AF subscale, and from .57 to .42 for the IR subscale.

INITIAL MULTIDIMENSIONAL ANALYSIS

So far, we had fit our data to the Rasch measurement model in consecutive analysis for the three OFER subscales. Despite the small size of our sample, we conducted an initial multidimensional analysis of the OFER scales using ConQuest software (Adams, Wu, & Wilson, 2015). We stress that our multidimensional analysis is exploratory and that we intend to further examine the results we report here when we have analyzed Lebanese nurses' responses to the national survey we conducted to investigate the relationship between the organization of nursing work in Lebanon, occupational fatigue, and muscular skeletal injury.

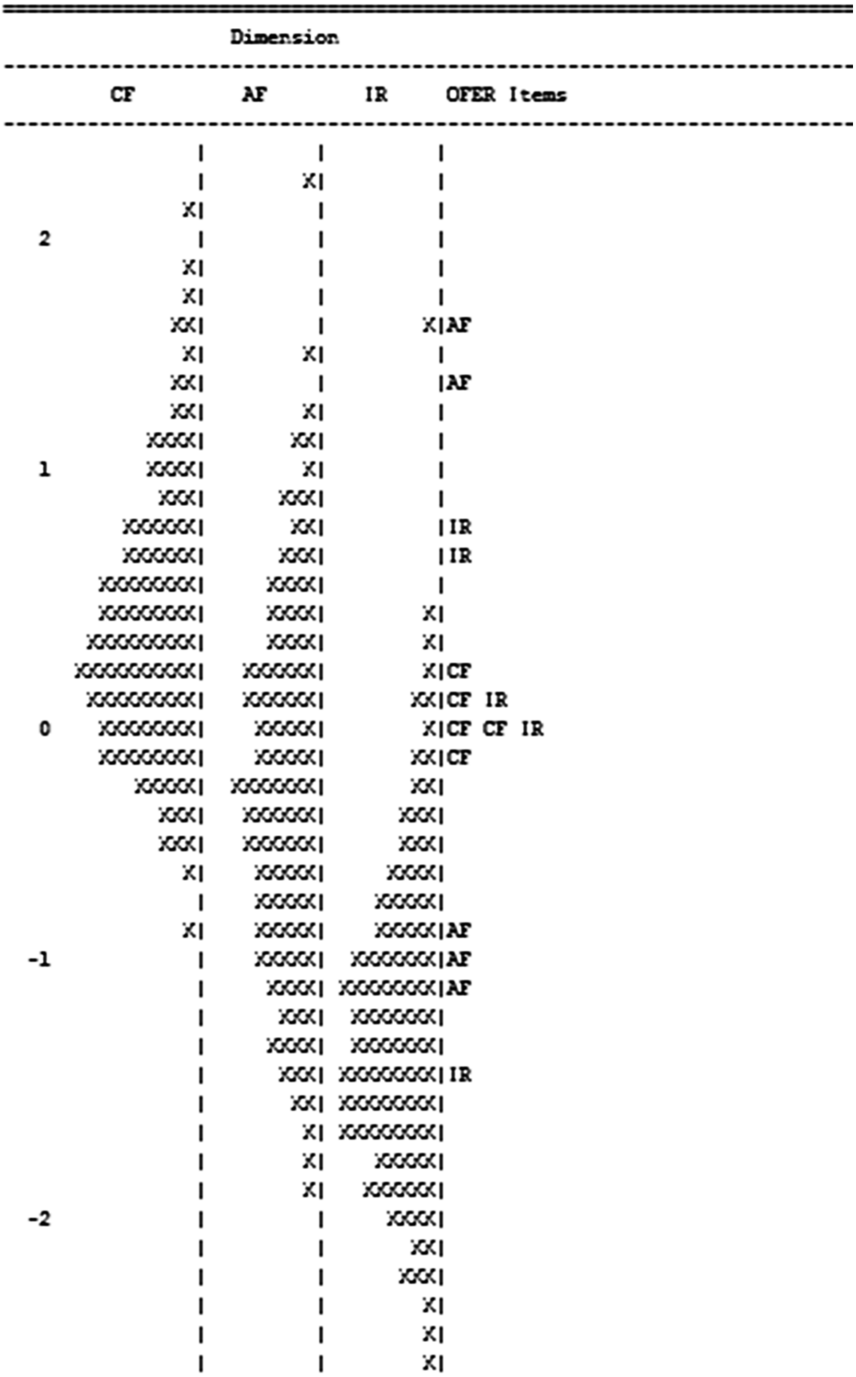
The difference in deviance statistics between nested models approximates a chi-squared distribution; degrees of freedom are given by the difference in the number of parameters estimated. The OFER subscale data were a better fit to the three-dimensional model: difference, $\chi^2 = 104$, $df = 12$, $p < .01$; unidimensional Akaike information criterion (AIC) = 11,925; and multidimensional AIC = 11,821. The relationship between the three OFER subscales was clearly not orthogonal. The disattenuated correlations between the subscales were .772 (CF and AF), .776 (CF and IR), and .707 (AF and IR). Reliability of the AF subscale increased from .63 to .73 when the data were fit to the three-dimensional model. However, the reliability of the CF subscale declined from .83 to .65 and that of the IR subscale from.57 to .48. These results confirm the acceptable reliability of the CF and AF subscales, the unacceptably low reliability of the IR subscale, and the need for further exploration of the fit of OFER items. We applied the Spearman–Brown formula and noted that six additional items are needed to increase the reliability of the CF subscale to .80 and that additional 17 items are required to achieve the same reliability for the IR subscale. Reliability of .80 for the AF subscale can be achieved with an additional three items.

If reliability of .70 is sufficient for the purposes of investigation, two items are required for the CF subscale and six for the IR subscale.

The multidimensional Wright's map in Figure 1 shows the distribution of respondent measures, mean item calibrations, item targeting, and the ceiling and floor effects of the OFER subscales. The gaps in the distribution of items relative to that of the respondents indicate where additional items are needed to improve subscale sensitivity and reliability.

NORM VALUES

We report unidimensional and multidimensional Rasch respondent measures and norm values for the OFER subscales in Table 3 to enable comparisons. The CF, AF, and IR scores reported in Table 3 are OFER subscale raw scores. The cutoff points for our sample are as follows: CF (low ≤28, low/moderate 29–≤52, moderate/high 53–≤76, high ≥77), AF (low ≤35, low/moderate 36–≤57, moderate/high 58–≤79, high ≥80), IR (low ≤25, low/moderate 26–≤50, moderate/high 51–≤75, high ≥76).



```
======================================================================
                  Dimension
----------------------------------------------------------------------
         CF         AF        IR    OFER Items
----------------------------------------------------------------------
                |         |         |
                |        X|         |
               X|         |         |
   2            |         |         |
               X|         |         |
               X|         |         |
              XX|         |        X|AF
               X|        X|         |
              XX|         |         |AF
              XX|        X|         |
            XXXX|       XX|         |
   1        XXXX|        X|         |
             XXX|      XXX|         |
          XXXXXX|       XX|         |IR
          XXXXXX|      XXX|         |IR
        XXXXXXXX|     XXXX|         |
        XXXXXXXX|     XXXX|        X|
       XXXXXXXXX|     XXXX|        X|
      XXXXXXXXXX|   XXXXXX|        X|CF
       XXXXXXXXX|   XXXXXX|       XX|CF IR
   0    XXXXXXXX|    XXXXX|        X|CF CF IR
        XXXXXXXX|    XXXXX|       XX|CF
           XXXXX|  XXXXXXX|       XX|
             XXX|   XXXXXX|      XXX|
             XXX|   XXXXXX|      XXX|
               X|    XXXXX|     XXXX|
                |    XXXXX|    XXXXX|
               X|    XXXXX|    XXXXX|AF
  -1            |    XXXXX|  XXXXXXX|AF
                |     XXXX| XXXXXXXX|AF
                |      XXX|  XXXXXXX|
                |     XXXX|  XXXXXXX|
                |      XXX| XXXXXXXX|IR
                |       XX| XXXXXXXX|
                |        X| XXXXXXXX|
                |        X|    XXXXX|
                |        X|   XXXXXX|
  -2            |         |     XXXX|
                |         |       XX|
                |         |      XXX|
                |         |        X|
                |         |        X|
                |         |        X|
```

Figure 1. Map of latent distributions and response model parameter estimates for Occupational Fatigue Exhaustion Recovery (OFER) subscales. Each *X* = 3.4 respondents; items identified by subscale only to comply with copyright requirements. CF = Chronic Fatigue; AF = Acute Fatigue; IR = Intershift Recovery.

TABLE 3. Unidimensional and Multidimensional Rasch Measures and Norm Values for Occupational Fatigue Exhaustion Recovery Subscales

CF Subscale				AF Subscale				IR Subscale			
CF Raw	Measure Uni.	Measure Multi.	CF Score	AF Raw	Measure Uni.	Measure Multi.	AF Score	IR Raw	Measure Uni.	Measure Multi.	IR Score
0	−5.19	−1.81	0	0	−4.19	−2.56	13	0	−4.22	−2.65	0
1	−3.86	−1.06	7	1	−2.93	−2.10	17	1	−2.96	−2.21	3
2	−3.00	−0.92	7	2	−2.17	−1.86	7	2	−2.21	−1.94	7
3	−2.44	−0.81	10	3	−1.73	−1.71	10	3	−1.76	−1.73	10
4	−2.03	−0.73	13	4	−1.43	−1.60	13	4	−1.45	−1.50	13
5	−1.73	−0.66	17	5	−1.21	−1.51	17	5	−1.22	−1.22	17
6	−1.49	−0.60	20	6	−1.04	−1.43	20	6	−1.03	−0.95	20
7	−1.29	−0.55	23	7	−0.90	−1.35	23	7	−0.88	−0.76	23
8	1.12	−0.49	27	8	−0.77	−1.28	27	8	−0.74	−0.63	27
9	−0.98	−0.44	30	9	−0.66	−1.20	30	9	−0.62	−0.51	30
10	−0.84	−0.40	33	10	−0.55	−1.12	33	10	−0.51	−0.42	33
11	−0.71	−0.35	37	11	−0.45	−1.03	37	11	−0.41	−0.33	37
12	−0.59	−0.30	40	12	−0.35	−0.92	40	12	−0.32	−0.24	40
13	−0.47	−0.26	43	13	−0.25	−0.79	43	13	−0.22	−0.16	43
14	−0.36	−0.21	47	14	−0.15	−0.61	47	14	−0.13	−0.08	47
15	−0.24	−0.16	50	15	−0.04	−0.35	50	15	−0.04	−0.00	50
16	−0.11	−0.11	53	16	0.07	−0.00	53	16	0.06	0.07	53
17	0.02	−0.05	57	17	0.18	0.36	57	17	0.15	0.15	57
18	0.16	0.00	60	18	0.29	0.61	60	18	0.25	0.23	60

19	0.32	0.07	63	19	0.41	0.78	63	19	0.35	0.31	63
20	0.50	0.14	66	20	0.53	0.91	66	20	0.46	0.39	66
21	0.70	0.24	70	21	0.65	1.02	70	21	0.58	0.47	70
22	0.95	0.35	73	22	0.78	1.13	73	22	0.71	0.56	73
23	1.24	0.49	77	23	0.91	1.23	77	23	0.85	0.67	77
24	1.58	0.67	80	24	1.07	1.35	80	24	1.03	0.79	80
25	1.98	0.92	83	25	1.24	1.48	83	25	1.23	0.94	83
26	2.44	1.26	87	26	1.47	1.64	87	26	1.49	1.16	87
27	2.97	1.72	90	27	1.77	1.89	90	27	1.84	1.47	90
28	3.59	2.48	93	28	2.22	2.31	93	28	2.31	1.93	93
29	4.49	3.73	97	29	2.99	3.10	97	29	2.70	3.10	97
30	5.82	—	100	30	4.26	4.39	100	30	3.96	4.37	100

Note. CF = Chronic Fatigue; AF = Acute Fatigue; IR = Intershift Recovery; Uni. = unidimensional; Multi. = multidimensional.

DISCUSSION

Rating Scale Categories

The OFER seven-category rating scale confused our sample. The categories *slightly disagree* and *slightly agree* were too nuanced for respondents whose first language is Arabic. A numerical rating scale with low and high anchor points and no descriptors might avoid similar problems in future studies. Another solution is to use a three-category rating scale. For example, *disagree*, *neutral*, and *agree*. This solution is suggested by our finding that none of the Rasch–Andrich thresholds for any of the three scales increased by at least 1.4 logits. However, reducing the number of rating scale categories would reduce raw score variance and measurement sensitivity. Therefore, we recommend that investigators provide OFER respondents with prior training in using seven-category rating scales.

Studies are required to determine whether rating category disordering, poor targeting, and the other problems with the OFER subscales remain when the scale is administered in Arabic. The OFER is available for administration in French. Therefore, a future study could examine the measurement performance of OFER subscales when administered in English, Arabic, and French. This would be feasible in Lebanon because all three languages are widely spoken. The Rasch–Andrich threshold disordering we identified is due to the narrow range of the measures on the latent variable. The solution is to recruit a larger and more diverse sample of respondents.

Wording and Language

Except for our findings on rating scale category disordering and negatively keyed items, there is no evidence that the respondents had difficulty with the wording of the OFER items. This implies that subject to prior permission from the copyright holders, cultural testing (Collins, 2014; Miller, Chepp, Willson, & Padilla, 2014), and reconsideration of negatively keyed items, the OFER subscales are suitable for translation into Arabic.

Dimensionality and Item Analysis

Our results confirm the unidimensionality of the CF subscale. Future studies with larger samples will provide opportunities to further investigate the dimensionality of the AF and IR subscales using sample size adjusted cutoff values (Smith et al., 1998). Studies with large enough samples will support exploratory and confirmatory factor analysis and comparison of unidimensional and plausible multidimensional models. According to Embretson and Reise (2000), at least 500 respondents will be required. Parameters estimated with smaller samples are unlikely to be reliable. Kose and Demirtasli (2012) advise that longer tests (20 or more items) and samples of at least 1,000 respondents are required to achieve small error estimates. Conversely, reasonably stable Rasch parameters can be estimated with samples of 50 (Linacre, 1994). Studies with larger samples will enable reexamination of item dependency and possible item bias (DIF).

Targeting

Analyses of targeting showed that the CF subscale was targeted to respondents close to the sample mean. The AF subscale was better targeted to the subscale respondents with high and low measures; none of the items matched respondents with measures at or near the mean of the distribution. Except for one item, the IR subscale targeted respondents with measures at and above the sample mean. All three subscales had significant ceiling and floor effects.

Respondent Separation and Respondent Reliability

The CF subscale was the only OFER subscale to return a respondent separation ratio higher than 2.0, the minimum required (Linacre, 2015a) to distinguish between respondents with low, medium, and high location measures. Consequently, the CF subscale was the only OFER measure to achieve a respondent reliability value higher than .80, the minimum required (Linacre, 2015a) for a reliable sale. The reliability of the AF and IR subscales could be improved by reversing the negatively keyed items or by adding more items. Our preferred strategy is to increase variance in the Rasch respondent measures by recruiting a more heterogeneous sample. Samples recruited from across the health care sector with a cross section of nurses will likely improve the measurement performance of the AF and IR subscales and extend their operational range.

LIMITATIONS

Our choice of the Rasch measurement model precluded us from examining item discrimination, which is automatically constrained to be equal for all items by the WINSTEPS computer program (Linacre, 2015b). WINSTEPS does estimate item discrimination parameters, but we have not reported them because our study is not an application of the two-parameter item response model. Similarly, we have not estimated upper and lower asymptotes for ICCs because they are only relevant to applications of three-parameter and four-parameter item response models.

Our sample may have been too homogeneous to establish the reliability of the AF and IR subscales. We recruited registered nurses at one academic medical center with high levels of proficiency in English who cannot be considered representative of the nursing workforce in Lebanon. Furthermore, in using the unidimensional model, we have oversimplified the interaction between OFER items and our sample because respondents use several traits rather than one (Reckase, 2009) when replying to scale items. These limitations will be avoided in future studies by recruiting a more diverse sample of nurses and by conducting future studies in Arabic and French as well as in English. A sample of at least 1,000 nurses with oversampling of respondents with high and low levels of occupational fatigue and nurses in rural areas will permit multidimensional Rasch analyses. Our national study will help us to identify the characteristics of nurses and practice settings associated with lower and higher levels of occupational fatigue.

CONCLUSION

We used consecutive unidimensional Rasch analyses and an exploratory multidimensional analysis to investigate the psychometric characteristics of the OFER subscales. The unidimensional Rasch rating scale model was a good fit to the CF data, but the respondent measures for the AF subscale were poorly targeted to this sample. The IR subscale was better targeted but had poor reliability. Our multidimensional analysis enabled simultaneous calibration of the three OFER subscales, increasing measurement precision by including an assessment of the correlations between subscales. Further examination of the results of our multidimensional analysis was precluded by the size of our sample. Additional items may be required to investigate occupational fatigue among respondents in the Eastern Mediterranean region. Larger studies with more diverse samples are needed to further examine the reliability of the OFER subscales and possible differential item function. The convergent and discriminant validity of all three subscales needs to be established when investigating the validity of an Arabic version of the OFER.

REFERENCES

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised item response modelling software (Version 4) [Computer software]. Victoria, Australia: Australian Council for Educational Research.

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

Barker, L. M., & Nussbaum, M. A. (2011). Fatigue, performance and the work environment: A survey of registered nurses. *Journal of Advanced Nursing*, *67*(6), 1370–1382. http://dx.doi.org/10.1111/j.1365-2648.2010.05597.x

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. New York, NY: Springer Publishing.

Brooks, I. (2000). Nurse retention: Moderating the ill-effects of shiftwork. *Human Resource Management Journal*, *10*(4), 16–31. http://dx.doi.org/10.1111/j.1748-8583.2000.tb00003.x

Chana, N., Kennedy, P., & Chessell, Z. J. (2015). Nursing staffs' emotional well-being and caring behaviours. *Journal of Clinical Nursing*, *24*(19–20), 2835–2848. http://dx.doi.org/10.1111/jocn.12891

Chau, Y. M., West, S., & Mapedzahama, V. (2014). Night work and the reproductive health of women: An integrated literature review. *Journal of Midwifery & Women's Health*, *59*(2), 113–126. http://dx.doi.org/10.1111/jmwh.12052

Chen, J., Davis, K. G., Daraiseh, N. M., Pan, W., & Davis, L. S. (2014). Fatigue and recovery in 12-hour dayshift hospital nurses. *Journal of Nursing Management*, *22*(5), 593–603. http://dx.doi.org/10.1111/jonm.12062

Choi, B., Dobson, M., Landsbergis, P., Ko, S., Yang, H., Schnall, P., & Baker, D. (2014). Job strain and obesity. *Journal of Internal Medicine*, *275*(4), 438–440. http://dx.doi.org/10.1111/joim.12173

Collins, D. (2014). *Cognitive interviewing practice*. Los Angeles, CA: Sage.

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

Eriksen, W. (2006). Work factors as predictors of persistent fatigue: A prospective study of nurses' aides. *Occupational and Environmental Medicine*, *63*(6), 428–434. http://dx.doi.org/10.1136/oem.2005.019729

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 1095.

Frantom, C. G., Green, K. E., & Hoffman, E. R. (2002). Measure development: The Children's Attitudes toward Technology Scale (CATS). *Journal of Educational Computing Research*, *26*(3), 249–263. http://dx.doi.org/10.2190/DWAF-8LEQ-74TN-BL37

Granger, C. (2008). Rasch analysis is important to understand and use for measurement. *Rasch Measurement Transactions*, *21*(3), 1122–1123.

Heikkilä, K., Fransson, E. I., Nyberg, S. T, Zins, M., Westerlund, H., Westerholm, P., . . . Kivimäki, M. (2013). Job strain and health-related lifestyle: Findings from an individual-participant meta-analysis of 118,000 working adults. *American Journal of Public Health*, *103*(11), 2090–2097.

Heikkilä, K., Nyberg, S. T., Fransson, E. I., Alfredsson, L., De Bacquer, D., Bjorner, J. B., . . . Kivimäki, M. (2012). Job strain and tobacco smoking: An individual-participant data meta-analysis of 166, 130 adults in 15 European studies. *PLoS One*, *7*(7), e35463.

Kivimäki, M., Nyberg, S. T., Batty, G. D., Fransson, E. I., Heikkilä, K., Alfredsson, L., . . . Theorell, T. (2012). Job strain as a risk factor for coronary heart disease: A collaborative meta-analysis of individual participant data. *Lancet*, *380*(9852), 1491–1497. http://dx.doi.org/10.1016/S0140-6736(12)60994-5

Kose, A., & Demirtasli, N. C. (2012). Comparison of unidimensional and multidimensional models based on item response theory in terms of both variables of test length and sample size. *Procedia - Social and Behavioral Sciences*, *46*,135–140. http://dx.doi.org/10.1016/j.sbspro.2012.05.082

Kunaviktikul, W., Wichaikhum, O., Nantsupawat, A., Nantsupawat, R., Chontawan, R., Klunklin, A., . . . Sirakamon, S. (2015). Nurses' extended work hours: Patient, nurse and organizational outcomes. *International Nursing Review*, *62*(3), 386–393. http://dx.doi.org/10.1111/inr.12195

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2015a). *A user's guide to Winsteps: Rasch-model computer programs*. Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2015b). WINSTEPS® (Version 3.92.0) [Computer software]. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com

Masters, G. N. (1982). Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Miller, K., Chepp, V., Willson, S., & Padilla, J. L. (Eds.). (2014). *Cognitive interviewing methodology*. Hoboken, NJ: Wiley.

Rasch, G. A. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150). New York, NY: Springer Publishing.

Rella, S., Winwood, P. C., & Lushington, K. (2009). When does nursing burnout begin? An investigation of the fatigue experience of Australian nursing students. *Journal of Nursing Management*, *17*(7), 886–897. http://dx.doi.org/10.1111/j.1365-2834.2008.00883.x

Rost, J. (2001). What is a Rasch model? In A. Boomsma, M. van Duijn, & T. Snijders (Eds.), *Essays on item response theory* (pp. 26–30). New York, NY: Springer Publishing.

Samra, H. A., & Smith, B. A. (2015). The effect of staff nurses' shift length and fatigue on patient safety and nurses' health: From the National Association of Neonatal Nurses. *Advances in Neonatal Care*, *15*(5), 311. http://dx.doi.org/10.1097/ANC.0000000000000230

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66–78.

Winwood, P. C., Lushington, K., & Winefield, A. H. (2006). Further development and validation of the Occupational Fatigue Exhaustion Recovery (OFER) scale. *Journal of Occupational and Environmental Medicine*, *48*(4), 381–389.

Winwood, P. C., Winefield, A. H., & Lushington, K. (2006). Work-related fatigue and recovery: The contribution of age, domestic responsibilities and shiftwork. *Journal of Advanced Nursing*, *56*(4), 438–449. http://dx.doi.org/10.1111/j.1365-2648.2006.04011.x

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

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

Correspondence regarding this article should be directed to Michael Clinton, PhD, RN, Hariri School of Nursing, American University of Beirut, PO Box 11-0236, Riad El-Solh, Beirut 1107 2020. E-mail: mc42@aub.edu.lb