International Journal of Behavioral Medicine, 15: 62–72, 2008

ISSN: 1070-5503 print / 1532-7558 online
DOI: 10.1080/10705500701783959

Routledge
Taylor & Francis Group

Factorial Invariance and Stability of the Effort-Reward Imbalance Scales: A Longitudinal Analysis of Two Samples with Different Time Lags

Jan de Jonge, Sjaak van der Linden, Wilmar Schaufeli, Richard Peter, and Johannes Siegrist

Background; *Key measures of Siegrist's (1996) Effort-Reward Imbalance (ERI) Model (i.e., efforts, rewards, and overcommitment) were psychometrically tested.* ***Purpose***: *To study change in organizational interventions, knowledge about the type of change underlying the instruments used is needed. Next to assessing baseline factorial validity and reliability, the factorial stability over time—known as alpha-beta-gamma change—of the ERI scales was examined.* ***Methods***: *Psychometrics were tested among 383 and 267 healthcare workers from two Dutch panel surveys with different time lags.* ***Results***: *Baseline results favored a five-factor model (i.e., efforts, esteem rewards, financial/career-related aspects, job security, and overcommitment) over and above a three-factor solution (i.e., efforts, composite rewards, and overcommitment). Considering changes as a whole, particularly the factor loadings of the three ERI scales were not equal over time. Findings suggest in general that moderate changes in the ERI factor structure did not affect the interpretation of mean changes over time.* ***Conclusion***: *Occupational health researchers utilizing the ERI scales can feel confident that self-reported changes are more likely to be due to factors other than structural change of the ERI scales over time, which has important implications for evaluating job stress and health interventions.*

Key words: Effort-Reward Imbalance, overcommitment, ERI-Q scales, panel survey, alpha-beta-gamma change

The Effort-Reward Imbalance (ERI) Model, originally formulated by Siegrist and colleagues (Siegrist, 1996; Siegrist, Siegrist, & Weber, 1986), has received considerable attention in occupational health research merely due to its predictive power for adverse health and well-being outcomes (cf., van Vegchel, de Jonge, Bosma, & Schaufeli, 2005). Moreover, albeit to a lesser extent, the model has been used in organizational intervention studies as well (cf., Tsutsumi & Kawakami, 2004). The ERI Model has its origin in medical sociology and emphasizes both the effort and the reward structure of work (Marmot, Siegrist, & Theorell, 2006). According to the model, work-related benefits depend upon a reciprocal relationship between efforts and rewards at work. Efforts represent job demands and/or obligations that are imposed on the employee, such as time pressure and working overtime. Occupational rewards distributed by the employer (and by society at large) consist of money, esteem, and job security/career opportunities. More specifically, the ERI Model claims that work characterized by both high efforts and low rewards represents a reciprocity deficit between high "costs" and low "gains," which could elicit negative emotions in exposed employees. The accompanying feelings may cause sustained strain reactions. So, working hard without receiving adequate appreciation or being treated fairly are examples of a stressful imbalance.

Another assumption of the ERI Model concerns individual differences in the experience of effort-reward imbalance. It is assumed that employees characterized

Jan de Jonge, Subdepartment of Human Performance Management, Eindhoven University of Technology, and Department of Social and Organizational Psychology, Utrecht University, The Netherlands; Sjaak van der Linden, Department of Social and Organizational Psychology, Utrecht University, The Netherlands, and Zorgcirkel Waterland, The Netherlands; Wilmar Schaufeli, Department of Social and Organizational Psychology, Utrecht University, The Netherlands; Richard Peter, Department of Epidemiology, University of Ulm, Germany; and Johannes Siegrist, Department of Medical Sociology, University of Düsseldorf, Germany.

We would like to thank Magnus Sverke for his statistical advice during the analytical process and the two anonymous reviewers of this manuscript for their very helpful comments.

Correspondence concerning this article should be addressed to Jan de Jonge, Subdepartment of Human Performance Management, Department of Technology Management, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. E-mail: j.d.jonge@tue.nl

by a motivational pattern of excessive job-related commitment and a high need for approval (i.e., overcommitment) will respond with more strain reactions to an effort-reward imbalance, in comparison with less overcommitted people. As there is some evidence of intrapersonal stability of overcommitment over time (cf., Siegrist, 1996), it can be considered a risk factor in its own, even when effort-reward imbalance is absent. However, the ERI Model posits that strongest adverse health and well-being effects take place if work and personal conditions act simultaneously (cf., Siegrist et al., 2004).

Measurement of Effort-Reward Imbalance

To measure the core aspects of the ERI Model, information was gathered from different sources; that is (1) contextual information such as administrative data, and (2) descriptive and evaluative information through interviews and self-report questionnaires. A combination of those sources was initially used to measure efforts and rewards, whereas overcommitment was assessed by a self-report questionnaire. Subsequently, a self-report questionnaire was developed to measure all key components of the ERI Model, i.e., efforts, rewards, and overcommitment. The introduction of this so-called ERI Questionnaire (ERI-Q; Siegrist & Peter, 1996) led to a predominant use of self-report questionnaires to test the ERI Model (e.g., see Siegrist et al., 2004).

The ERI-Q encompasses the three key concepts of the ERI Model, i.e., efforts, rewards, and overcommitment (Siegrist & Peter, 1996). The ERI-Q is restricted to self-report data because it combines descriptive and evaluative information on *perceived* demands (efforts) and rewards. Theoretically, three components underlie the concept of rewards, distinguishing salary, esteem, and security/career opportunities (Siegrist et al., 2004). Next to assuming that the three factors load on one latent factor ("composite rewards"; cf., Rödel, Siegrist, Hessel, & Brähler, 2004), it was postulated that a three-factor structure would fit the data as well. In addition, the ERI-Q requires information on the *personal* characteristic "overcommitment."

Assessment of efforts, rewards, and overcommitment relies on indicators that are measured by psychometric scales containing Likert-scaled items (cf., Siegrist et al., 2004). Effort and reward items are answered in two separate steps. First, subjects indicate whether a given type of (stressful) condition at work exists by choosing between two response categories: "agree" versus "disagree." If the respondent "agrees," he or she is asked to indicate the degree of distress of this condition on a 4-point scale ranging from 1 (not at all distressed) to 4 (very distressed). A negative answer (i.e., disagree) is also coded as 1. A sum score of these ratings is constructed accordingly. Empirical studies showed psychometrically appropriate versions of efforts, rewards, and overcommitment, respectively. For instance, the uni-dimensionality of the effort scale has been documented in several studies (e.g., Hanson, Schaufeli, Vrijkotte, Plomp, & Godaert, 2000; Peter et al., 1998; Tsutsumi, Ishitake, Peter, Siegrist, & Matoba, 2001; Siegrist et al., 2004). Furthermore, studies confirm the presence of a uni-dimensional reward structure (e.g., Joksimovic, Starke, von dem Knesebeck, & Siegrist, 2002; Siegrist et al., 2004) as well as a three-factorial structure of rewards (e.g., Dragano, von dem Knesebeck, Rödel, & Siegrist, 2003; van Vegchel, de Jonge, Bakker, & Schaufeli, 2002; Siegrist et al., 2004). Finally, as far as overcommitment is concerned, studies reported acceptable psychometric properties (e.g., Joksimovic et al., 2002; Siegrist et al., 2004).

The Present Study

The aim of this study is to further investigate the validity and reliability of the Effort-Reward Imbalance Questionnaire (ERI-Q). Earlier research provides a good foundation for applying a confirmatory approach in investigating the ERI-Q. As stated previously, there are also organizational intervention studies based upon the ERI Model. To capture organizational change, researchers have relied primarily on comparing (self-report) measures given at entry and again at some later point/s. To ascertain whether an organizational intervention has succeeded or failed requires not only measuring the quantity of change, but particularly demands confidence in the concept of change that underlies the measurement (Golembiewski, Billingsley, & Yeager, 1976). So, the crucial question is whether it is the employees who have changed or the tests? Interpreting any results of intervention research is chancy in the absence of knowledge about types of change. Three types of change can be defined in summary fashion, as we shall further distinguish them in the analysis (cf., Golembiewski et al., 1976). A first type of change is *alpha change*: the change in the level of a variable from one measurement point to the next given a constantly calibrated instrument related to a constant conceptual domain. The second type, called *beta change*, occurs when respondents recalibrate the measurement continuum (e.g., a score 5 at Time 2 may be defined as was 4 at Time 1). The last type of change is called *gamma change* and involves a redefinition of the construct underlying the instrument. One can imagine that any comparisons between time points are meaningless under gamma change because the instrument now operationalizes different constructs at both points. So, demonstration of time invariance of the ERI-Q scales allows researchers to conduct organizational intervention studies without having to worry about whether observed differences and real changes in efforts, rewards, and overcommitment are due to structural change of the ERI-Q constructs over time. To date, the stability of ERI-Q constructs and operationalizations over time have not been investigated, which has important implications for both research and practice. A last point of attention is the appropriateness of the time lag used for organizational change. As Frese and Zapf (1988) have noted, there is little information available about the "right" length of time lags in occupational health research. Ideally, the time lag of a research study encompasses the potential true change in the organization. One way to provide more information about appropriate time lags in longitudinal studies is to examine as many different lags as possible (cf., Dormann & Zapf, 2002).

For that very reason, we will examine the psychometric properties of the ERI-Q in several ways. First, baseline factorial validity and reliability of the ERI-Q will be assessed. Second, we will investigate the stability of ERI-Q constructs and operationalizations over time (i.e., the issue of alpha-beta-gamma change). Finally, we will use two panel samples with different time lags (i.e., a one-year and a two-year time interval, respectively) to test whether or not different time intervals influence the results accordingly.

Method

Study Samples

This paper is based on data obtained from two panel samples consisting of all healthcare employees working in two large organizations for residential elderly care in the Netherlands. Questionnaires contained an administration number for second-round identification (which was only known to the researchers) and could be returned in sealed envelopes. Both studies consist of a full panel design with two panel waves: subjects supplied data at two points in time (with nearly identical starting dates) with a one-year and two-year time interval, respectively. These time lags appear to be long enough for possible changes in individual scores, but not too long for a too high rate of attrition in the study sample. In addition, in this way possible seasonal fluctuations in work were controlled. To determine whether attrition might have biased results, we used logistic regression analyses to test whether participation at Time 2 was related to any Time 1 variable (cf., Goodman & Blum, 1996). These two analyses produced no significant terms, indicating that the attrition was random. Finally, it should be noted that there was no structured, planned intervention in both studies. Only natural and minor organizational changes took place, which had to do with some organizational renewal and personnel changes between the two waves.

All employees working at the organizations were included in the study (n = 894 for Study 1 and n = 554 for Study 2). With respect to Study 1, 894 employees received the questionnaire at Time 1, and the response rate was 66.1% (or 591 persons). Demographics showed that 89.2% of the sample was female, and mean age was 40.0 years (SD = 10.1, range 17–64). Mean working time was 7.8 years (SD = 7.0). At Time 2, one year later, 549 out of 902 people (or 60.9%) responded to the questionnaire. Note that Time 2 questionnaires went out to everyone in the sample, regardless of whether or not they completed and returned a Time 1 questionnaire. The panel group (those who responded to the questionnaire at both Time 1 and Time 2) consisted of 383 respondents (i.e., 64.8% of the initial group). Demographic characteristics of the respondents in the final panel of the first study showed that the ages ranged from 19–63 years (M = 40.3, SD = 9.7). Most respondents were female: 89.6%, and mean working time was 7.6 years (SD = 6.9).

As to Study 2, 405 out of 554 employees filled out the self-report questionnaires at Time 1 (73.1% response rate). Demographics showed that 90.8% of the respondents were female, and mean age was 38.8 years (SD = 8.7, range 16–62). Mean working time was 8.7 years (SD = 7.0). At Time 2, two years later, 420 out of 624 persons participated (67.3% response rate). The final panel consisted of 267 persons, or 48.2% of the initial group. Most of these respondents were female (91.4%). Ages ranged from 18–64 years (M = 41.0, SD = 8.7), and mean working time was 11.3 years (SD = 7.5).

Measures

A Dutch translation of the original, German ERI Questionnaire (ERI-Q; Siegrist & Peter, 1996) came into being by means of translation and independent back-translation to both the German and English versions of the instrument (see also Hanson et al., 2000).

Efforts were measured by six items with a 4-point scale ranging from 1 (not at all distressed) to 4 (very distressed). The content varies from physical load, time pressure, interruptions, responsibility, working overtime, to increasing demands. Example items are: "I have constant time pressure due to a heavy workload," and "I am often pressured to work overtime." Internal consistency of the effort scale, expressed by Cronbach's alpha, was satisfactory in both samples (both $\alpha = 0.74$).

Rewards were assessed by 11 items with a 4-point rating scale ranging from 1 (not at all distressed) to 4 (very distressed), which were reverse coded afterward. This scale was composed of three components, i.e., esteem rewards (5 items, e.g., "I receive the respect from my superiors") financial and career-related aspects (4 items, e.g., "My job promotion prospects are poor"), and job security (2 items, e.g., "My job security is poor"). Cronbach's alpha of the composite reward scale (11 items) was good in both samples (both $\alpha = 0.82$).

Overcommitment was measured by six items with rating scales ranging from 1 (strongly disagree) to 4 (strongly agree). The items consist of inability to withdraw from work and of impatience and disproportionate irritability. Example items are: "Work rarely lets me go, it is still on my mind when I go to bed," and "As soon as I get up in the morning I start thinking about work problems." Cronbach's alpha of this scale was satisfactory in Sample 1 ($\alpha_1 = 0.73$) and also in Sample 2 ($\alpha_2 = 0.78$).

Analytical Procedure

It should first be noted that all analyses were based on listwise deletion of missing data. We estimated cross-sectional confirmatory factor analytical models using LISREL 8.30 (cf., Jöreskog & Sörbom, 1996). First, a one-factor model was estimated proposing that all 23 items load on the same underlying dimension. Second, a model was estimated positing the original three factors representing efforts, rewards, and overcommitment. Third, a five-factor model was estimated to test the three-dimensional structure of rewards in addition to the theoretical three-factor model. Model tests were based upon the covariance matrix[1] and used maximum likelihood estimation. Model fit was assessed by a chi-square test with a non-significant test indicating a good fit to the empirical data. However, because non-significant chi-square test values are rarely obtained in this kind of analysis, we also used other fit indices such as adjusted goodness of fit index (AGFI), the non-normed fit index (NNFI), the parsimonious normed fit index (PNFI), the comparative fit index (CFI), and the root mean squared error of approximation (RMSEA). As the models estimated stand in a nested sequence, the relative fit of the models was tested through use of the chi-square difference test ($\Delta\chi^2$; Bentler & Bonett, 1980).

The analytical procedure for assessing mean change over time consisted of four phases (cf., Vandenberg & Self, 1993). First of all, we tested the equality of the variance-covariance matrices across time using LISREL. The purpose of this test was to provide an overall index in that rejecting the null hypothesis (i.e., the matrices are not equal over time) argues for testing of more restrictive hypotheses to identify the source of inequality. The second phase consisted of hierarchical tests of four models as defined by Schaubroeck and Green (1989), testing for the presence of gamma and beta change. As stated before, gamma change involves a redefinition of the construct under study, whereas beta change occurs when the respondents recalibrate the intervals anchoring the measurement continuum. For that reason, Model 1 examined changes in the number of factors over time, also known as the first test of gamma change. Model 2 extended this first test for gamma change by testing a model in which the factor covariances were constrained to be equal (i.e., the second test of gamma change). Model 3 examined a similar model as Model 2 but constraining the factor variances to be equal, which is called the first test of beta change. Finally, Model 4 investigated beta change at an item level (i.e., the second test of beta change). This model specifies the same patterns as Model 3, but places equality constraints on the factor loadings across time.

The third phase consisted of testing for alpha change: a change in the level of a variable over time given a constantly calibrated instrument related to a constant conceptual domain (Golembiewski et al., 1976). This was done by using repeated-measures multivariate analysis of variance (MANOVA) and paired t-tests to test for mean differences in the variables. The final phase examined the influence of gamma and beta change presence on alpha change, using LISREL again. In this case, the moment matrix associated with each construct was entered into the analysis (Bollen, 1989; see also Vandenberg & Self, 1993). Two models were estimated accordingly. In the first model, latent means were freely estimated, whereas in the second model, all latent means were constrained to be equal. Then the models were compared by means of a chi-square difference test, in which a significant deterioration in fit resulted in the rejection of the hypothesis of equal latent means. This test is analogous to the omnibus F-test of the MANOVA.

Finally, we analysed two panel samples with different time lags (one-year and two-year time intervals, respectively), which allowed for some variation in factorial invariance and stability testing.

Results

Table 1 summarizes the results of the baseline confirmatory factor analyses for the ERI-Q scales. Clearly, the one-factor model did not account well for the data in both samples. In addition, this table shows that the five-factor model provided the best fit to the data in both samples. More specifically, the five-factor model with three separate dimensions for reward provided a significantly better fit to the data than the three-factor model in both Sample 1 ($\Delta\chi^2(7) = 149.10$, $p < 0.001$) and Sample 2 ($\Delta\chi^2(7) = 131.76$, $p < 0.001$). All item loadings and factor correlations in the five-factor solution were significant ($p < 0.05$). To facilitate improvement of model fit (mainly in Sample 1), LISREL provides for each fixed parameter what is called a "Modification Index" (MI) that shows how much the model fit will improve if a parameter which was fixed to a specific value a priori (e.g., zero) is set free and is estimated

[1]Covariance matrices of both samples are available from the first author upon request.

Table 1. *Fit Indices for the Baseline Confirmatory Factor Analyses of the ERI-Q.*

Model	χ^2	*df*	AGFI	NNFI	PNFI	CFI	RMSEA
Sample 1 ($n = 368$)							
1-Factor	1309.97[a]	230	.64	.49	.44	.53	.14
3-Factor	640.41[a]	227	.83	.80	.67	.82	.07
5-Factor	491.31[a]	220	.86	.86	.70	.88	.06
5-Factor Respecified	422.80[a]	216	.88	.90	.71	.91	.05
Sample 2 ($n = 369$)							
1-Factor	931.23[a]	230	.61	.46	.41	.51	.14
3-Factor	479.93[a]	227	.80	.80	.64	.82	.07
5-Factor	348.17[a]	220	.85	.90	.69	.91	.05
5-Factor Respecified	304.87[a]	216	.87	.93	.70	.94	.04

[a] $p < 0.001$.

from the data. A closer inspection of the MIs showed significantly correlated error terms, which might be due to item similarity or common causes influencing the responses to these items that are not accounted for by the latent variables of the model. Stepwise relaxing the four corresponding parameters led to respecified models and improved model fit. The model modifications followed a consistent modification pattern in both samples and are quite defensible (see Figure 1). The within-factor correlated errors of both efforts (i.e., E2#E5, E2#E6, and E4#E5) and overcommitment (i.e., OC2#OC4) may reflect small group factors or item similarity (cf., Byrne, 1989). Both respecified five-factor models indicate a good fit in both samples, in which most of the fit indices reached common thresholds. Because a better fitting model is always obtained when more parameters are estimated, it is important to note that the respecified five-factor model also provided the best parsimonious fit to the data in both samples (PNFI = 0.71 and 0.70, respectively).

We will proceed with the results of the stability of constructs and operationalizations over time. As the five-factor model appeared to be the best fitting factorial model, this model will be the starting point for further analysis. First of all, test-retest coefficients showed a relatively stable effort construct over time in Sample 1 ($r = 0.58$, $p < 0.001$, one-year time lag). In Sample 2, however, effort was somewhat less stable ($r = 0.45$, $p < 0.001$, two-year time lag) given a rough threshold of 0.50 for this kind of constructs (cf., Bollen, 1989). In addition, test-retest coefficients for esteem rewards were $r = 0.24$ and $r = 0.30$ (both $p < 0.001$) for both samples, respectively, which is not very stable across time. For financial and career-related aspects, test-retest coefficients were $r = 0.34$ and $r = 0.41$ (both $p < 0.001$), respectively, whereas for job security, test-retest coefficients were $r = 0.31$ ($p < 0.001$) and $r = 0.10$ ($p =$ ns), respectively. Test-retests coefficients were $r = 0.37$ (Sample 1, $p < 0.001$) and $r = 0.38$ (Sample 2, $p < 0.001$) for the composite measure of rewards, which is not very stable across time. As far as overcommitment is concerned, test-retest coefficients showed a relatively stable construct in Sample 1 ($r = 0.53$, $p < 0.001$), whereas overcommitment was somewhat less stable in Sample 2 ($r = 0.45$, $p < 0.001$).

Further, as mentioned before, the analytical procedure for assessing mean change over time (i.e., alpha-beta-gamma change) consisted of four phases. Phase 1 of the analyses tested the equality of the variance-covariance matrix between the two waves. This initial analyses revealed that the chi-square value of the Sample 1 omnibus test was statistically significant ($\chi^2(276) = 841.94$, $p < 0.001$). In addition, the same appeared to be true for the chi-square value of the Sample 2 omnibus test: $\chi^2(276) = 515.39$, $p < 0.001$. So it would have to be generally concluded that the variance-covariance matrices were unequivalent between the two waves in both samples, which might be indicative of unstable measurement continua underlying the ERI-Q.

Results from the tests of more restrictive models (i.e., phase 2) are presented in Table 2. In Model 1 of this table, we examined the extent to which the five-factor model holds in both time periods (gamma 1 change). Initially, the values of the RMSEA (0.05 and 0.04 in Sample 1 and 2, respectively), and to a lesser extent the values of the NNFI (0.83 and 0.86, respectively) and CFI (0.85 and 0.87, respectively) indicated that not much gamma 1 change was present: a five-factor structure seemed to hold satisfactory across time. It should be stressed that this kind of model contains a huge number of measured variables and parameter estimates, implying that cut-off values of 0.90 on common goodness-of-fit measures such as the NNFI and CFI are not realistic (cf., Hair, Black, Babin, Anderson, & Tatham, 2006). The RMSEA is a more reliable indicator in this respect. However, the evidence through the

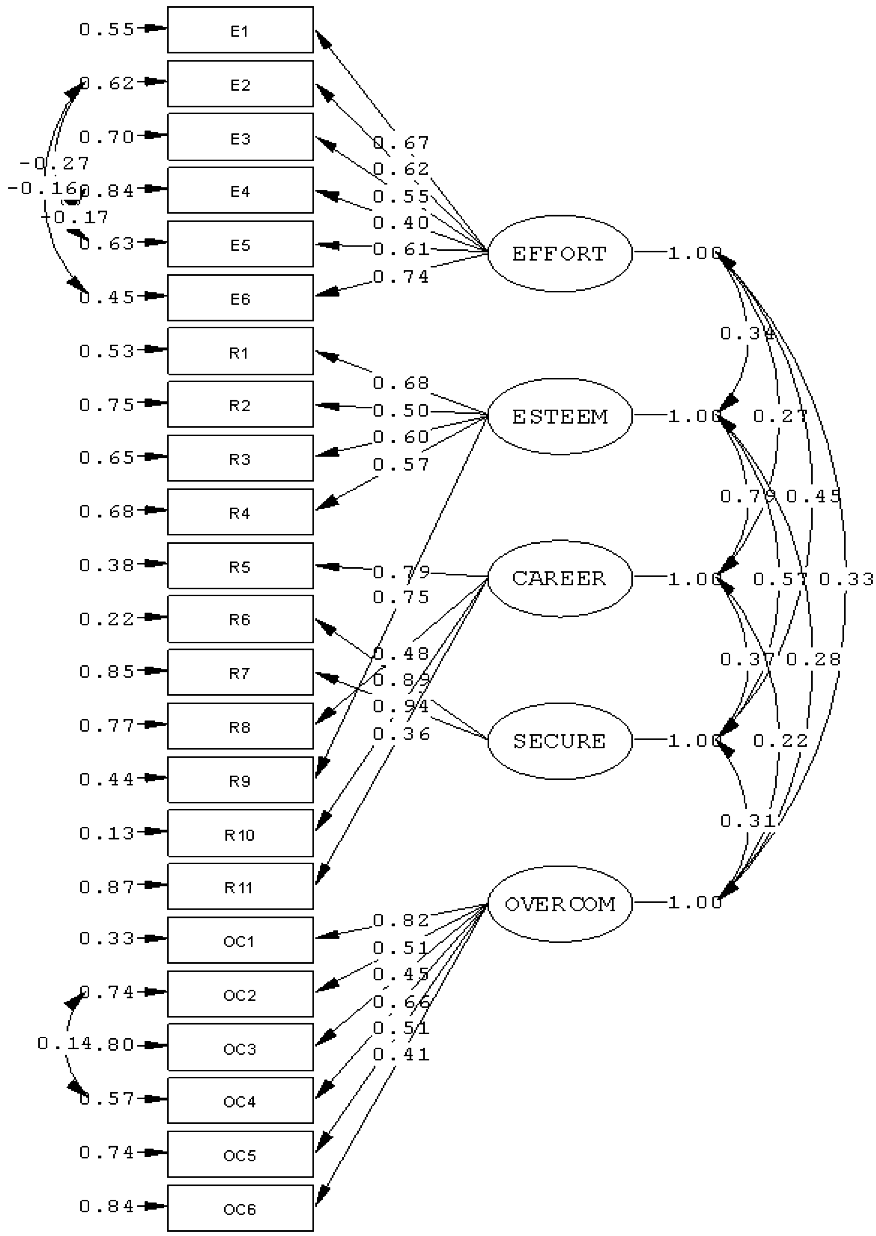

Figure 1. Five-factor solution (respecified) of ERI-Q confirmatory factor model (Sample 1; completely standardized solution; all loadings shown were significant at $p < 0.05$).

significant chi-square values in both samples indicated the *potential* presence of gamma and beta changes. Therefore, in Model 2 an additional type of gamma change was tested, i.e., the equivalence of factor covariances. The corresponding equality constraints in the models provided a non-significant chi-square difference test in Sample 1 ($\Delta\chi^2(10) = 14.43$, $p =$ ns) as well as in Sample 2 ($\Delta\chi^2(10) = 15.26$, $p =$ ns). This implies that the hypothesis of equal factor covariances should *not* be rejected (i.e., there is no evidence for the presence of gamma 2 changes across a one-year as well as a two-year time interval). In Model 3 the factor variances were specified to be equal across time (beta 1 change). The corresponding chi-square difference test was significant in Sample 1 ($\Delta\chi^2(5) = 39.06$, $p < 0.001$), indicating statistical evidence for the presence of beta 1 changes in Sample 1 (one-year time interval). Contrarily, the chi-square difference test was not significant in Sample 2: $\Delta\chi^2(5) = 5.03$, $p =$ ns), indicating no evidence for the presence of equal factor variances across a two-year time-interval. Finally, another form of change would involve the factor loadings (Model 4). Because there were 23 items under study and 5 of these served as a metric for a latent variable, the test statistic comparing Model 3 and Model 4 has 18 degrees of freedom. The chi-square difference test was significant in both samples, implying that the hypothesis of equal factor loadings across a one-year and

Table 2. *LISREL Tests for Gamma- and Beta-Change Presence in the Five-Factor Model (Samples 1 and 2; n = 318 and n = 231)*

Study Group	χ^2	*df*	NNFI	CFI	RMSEA	$\Delta\chi^2(\Delta df)$
Model 1 (Gamma 1): Same Factor Model Across Time						
Sample 1	1710.88[a]	921	.83	.85	.05	
Sample 2	1362.84[a]	921	.86	.87	.04	
Model 2 (Gamma 2): Equal Factor Covariances						
Sample 1	1725.31[a]	931	.83	.85	.05	14.43 (10)
Sample 2	1378.10[a]	931	.86	.87	.04	15.26 (10)
Model 3 (Beta 1): Equal Factor Variances						
Sample 1	1764.37[a]	936	.83	.84	.05	39.06[a] (5)
Sample 2	1383.13[a]	936	.86	.87	.04	5.03 (5)
Model 4 (Beta 2): Equal Factor Loadings						
Sample 1	1926.51[a]	954	.80	.82	.06	162.14[a] (18)
Sample 2	1473.39[a]	954	.84	.85	.04	90.26[a] (18)

[a] $p < 0.001$.

two-year time interval should be rejected (i.e., evidence for beta 2 changes in both samples).

Phase 3 of the analysis was the test for alpha change using a repeated-measures MANOVA. The corresponding results are presented in Table 3. Both repeated-measures MANOVAs revealed evidence of alpha change: Wilk's $F(5, 298) = 47.33$, $p < 0.001$ in Sample 1, and Wilk's $F(5, 217) = 6.30$, $p < 0.001$ in Sample 2. In-depth examination of independent t-tests showed that efforts increased significantly from Time 1 to Time 2 in Sample 1 ($t(302) = -11.83$, $p < 0.001$) and decreased significantly over time in Sample 2 ($t(221) = 4.05$, $p < 0.001$). Second, esteem rewards decreased significantly from Time 1 to Time 2 in Sample 1 ($t(302) = 10.66$, $p < 0.001$) and increased significantly over time in Sample 2 ($t(221) = -3.16$, $p < 0.01$). Third, financial and career-related aspects decreased significantly over time in Sample 1 ($t(302) = 9.56$, $p < 0.001$) and increased significantly over time in Sample 2 like esteem rewards ($t(221) = -4.12$, $p < 0.001$). Fourth, job security decreased significantly from Time 1 to Time 2 in Sample 1 ($t(302) = 6.06$, $p < 0.001$) and increased significantly over time in Sample 2 ($t(221) = -3.07$, $p < 0.01$). Finally, overcommitment decreased significantly over time only in Sample 2 ($t(221) = 2.52$, $p < 0.01$), whereas Sample 1 did not show any significant changes in overcommitment ($t(302) = -0.71$, $p =$ ns).

Evidence of alpha change was also present when LISREL modelling was applied (phase 4). These results are depicted in Table 4 for both Sample 1 and Sample 2. As can be seen, constraining the latent means for the variables to be equal resulted in a significant worsening of fit as evidenced by the differences in chi-square ($\Delta\chi^2$) in both Sample 1 ($\Delta\chi^2(5) = 94.81$, $p < 0.001$) and Sample 2 ($\Delta\chi^2(5) = 62.01$, $p < 0.001$). Furthermore, the pattern of differences between latent means did not change from the pattern associated with the MANOVAs. So the present findings suggest that moderate changes in factor structure did *not* affect the interpretation of mean differences with repeated measures in both samples.

Table 3. *Results from the Repeated Measures MANOVA*

Sample 1 ($n = 303$): Wilk's $F(5, 298) = 47.33$, $p < 0.001$					
	Means (SD)				
Time	**Efforts**	**Esteem**	**Career**	**Security**	**Overcommitment**
1	11.40 (3.52)	18.65 (2.37)	14.01 (2.43)	7.33 (1.19)	12.44 (3.12)
2	13.69 (3.83)	16.57 (3.06)	12.35 (2.83)	6.78 (1.47)	12.56 (2.90)
Sample 2 ($n = 222$): Wilk's $F(5, 217) = 6.30$, $p < 0.001$					
	Means (SD)				
Time	**Efforts**	**Esteem**	**Career**	**Security**	**Overcommitment**
1	11.89 (3.41)	18.42 (2.32)	13.71 (2.41)	7.29 (.99)	12.67 (2.89)
2	10.98 (2.95)	18.97 (2.00)	14.40 (2.14)	7.57 (1.06)	12.15 (2.94)

Table 4. *LISREL Tests for Effects of Gamma and Beta Change on Alpha Change*

Scales	χ^2	*df*	NNFI	CFI	RMSEA	$\Delta\chi^2$	Δdf
Sample 1 ($n = 318$)							
Model 1: Unequal Latent Means							
ERI-Q (5)	2493.60[a]	972	.95	.95	.07		
Model 2: Equal Latent Means							
ERI-Q (5)	2588.41[a]	977	.94	.95	.07	94.81[a]	5
Latent Means (Model 1)							
	Efforts	**Esteem**	**Career**	**Security**	**OC**		
Time 1	13.14	18.50	14.04	6.88	14.88		
Time 2	15.18	16.45	12.64	6.26	14.82		
Difference	2.04	−2.05	−1.39	−.62	−.06		
Sample 2 ($n = 231$)							
Scales	χ^2	*df*	**NNFI**	**CFI**	**RMSEA**	$\Delta\chi^2$	Δdf
Model 1: Unequal Latent Means							
ERI-Q (5)	2023.60[a]	972	.95	.95	.07		
Model 2: Equal Latent Means							
ERI-Q (5)	2085.61[a]	977	.95	.95	.07	62.01[a]	5
Latent Means (Model 1)							
	Efforts	**Esteem**	**Career**	**Security**	**OC**		
Time 1	13.86	17.90	13.44	6.86	14.04		
Time 2	11.94	18.75	14.44	7.22	13.32		
Difference	−1.92	.80	1.00	.36	−.72		

[a] $p < 0.001$.

Discussion

The objectives of the present study were to investigate the psychometric properties of a questionnaire measuring psychosocial stress at work in terms of the Effort-Reward Imbalance (ERI) Model (Siegrist, 1996). There is mounting evidence that researchers should no longer treat stable measurement instruments as a given, particularly in organizational intervention research (Vandenberg & Self, 1993). For that reason, we investigated factorial invariance across time and also the stability of and mean changes in the respective scales (i.e., alpha-beta-gamma change). To our knowledge, this is the first report assessing the stability of ERI constructs and operationalizations over time. Moreover, we used two panel samples with different time lags (one-year and two-year time intervals, respectively) to allow some variation in factorial invariance and stability testing.

For both baseline (i.e., Time 1) study samples, the factorial validity and reliability of the ERI-Q were assessed. The baseline factorial validity of the scales measuring the key components was replicated in a satisfactory way, of which a five-factor solution consisting of efforts, esteem rewards, financial/career-related aspects, job security, and overcommitment showed the best fit in both samples. In addition, the proposed factor model modifications followed a consistent modification pattern in both samples, and the corresponding correlated errors may reflect small group factors or item similarity in the effort and overcommitment scales. However, the correlated errors found are in contrast to the study of Hanson et al. (2000), who only found shared error variance in job security. Generally, several other ERI-Q studies found this five-factor structure, too (Hanson et al., 2000; Rödel et al., 2004). Recently, Rödel et al. (2004) showed that a second-order factor analysis revealed a higher-order general rewards factor as well. Future research should investigate whether such a higher-order factor may increase the factorial validity of the ERI-Q rewards scales. The internal consistencies of the scales, reflected by Cronbach's alpha, were ranging from "acceptable" (efforts, overcommitment) to "good" (rewards) in both samples. These results were identical to past ERI-Q research that also reported higher Cronbach's alphas for the rewards scale than for the other two scales (e.g., Hanson et al., 2000; Siegrist et al., 2004; Tsutsumi et al., 2001).

Analytical results of the stability of constructs and operationalizations over time can be split up in (1) test-retest analyses, and (2) testing for alpha-beta-gamma change. First, the test-retest coefficients showed not very stable constructs over time in both samples. Only efforts and overcommitment in Sample 1 reached an acceptable threshold of 0.50 for this kind of constructs (cf., Bollen, 1989). The coefficients for both efforts and overcommitment were somewhat lower in Sample 2, which might be caused by the longer time lag.

On the whole, test-retest coefficients tend to be higher for short-term retests than for long-term retests. The reason for this is that the true score may change over time (Bollen, 1989). In addition, stability coefficients are lowest for the rewards components in both samples. This might indicate that occupational rewards are more sensitive to change over time than efforts and even more stable characteristics like overcommitment. Many Dutch organizations do reorganize their organizational structure often (due to economic and labor market fluctuations), which might cause changes in job security and workplace social support. We know that employees from Sample 1 reported frictions due to organizational renewal and personnel changes between the two waves, which could explain the current findings. On the whole, the question rises whether or not we can expect ERI to be stable in terms of test-retest analysis. As ERI-Q tries to measure the employees' perception of the job situation, the scales should be sensitive to changes in job situations. Dynamic responsiveness of the ERI-Q to a series of organizational changes was indeed empirically shown by Tsutsumi, Nagami, Morimoto, and Matoba (2002). The measures declined for employees with stressful experiences due to reorganization, whereas they improved for employees who were promoted. So we could only expect ERI to be stable under non-changing job situations, which is seldom the case.

Further, as demonstrated in this article, one can have a much richer view of the changes in ERI-Q constructs over time through using a structural equations approach for testing alpha-beta-gamma change. More specifically, we moved to a longitudinal model that simultaneously took into account factor covariances over time and permitted the researchers to compare factor intercorrelations from time to time. The current results did not indicate a *potential* presence of gamma 1 changes in both samples: initially a five-factor structure seemed to hold across time. However, in analyzing alpha-beta-gamma change, it should be recognized that interpreting the baseline findings is not clearcut (cf., Vandenberg & Self, 1993). If focus is placed on significant chi-square values, then the evidence favored the existence of gamma 1 change in Model 1. If focus is placed on other, more reliable, fit indices (such as RMSEA), then this evidence is not convincing at all. Further simulation research is needed to determine if this approach is effective in detecting true changes in factor structure across time. Considering gamma and beta changes as a whole, the presence of any change was particularly pronounced for beta 2 change in both samples: the factor loadings of the three ERI-Q scales were not equal over time. This implies that the scaling unit for the respective items is not constant over time (i.e., change in metric for the constructs perceived) in both a one-year and a two-year condition. This might have some consequences for the response options in the ERI-Q. First of all, employees were asked to rate how "distressful" certain effortful and rewarding aspects of their work are. We know from work-psychological research that this way of rating merely reflects subjective and momentaneous feelings of distress, rather than a reflection of an objective job situation (e.g., Frese & Zapf, 1988). In this respect, Frese and Zapf proposed frequency rating scores instead of perceived distress. So it seems plausible that the rating of the ERI-Q items is more susceptible to individual variation, particularly over time. Further, as Tsutsumi (2004) has noted, the ERI-Q two-step format and coding used here is probably not the most valid one. To prevent misclassification and unanswered items, Tsutsumi (2004) has suggested using a response format with one step instead of two steps. Most ERI researchers have adopted this suggestion, and the most recent version of the ERI-Q was changed accordingly (cf., Siegrist et al., 2004). Future research is highly recommended to test for beta 2 changes in the latest ERI-Q version. Moreover, it is recommended that next to—or maybe even instead of—rating employees' level of distress (intensity), participants should rate *how often* they experience effortful and rewarding aspects as well (frequency). In practice, however, it appears that these two ways of rating are closely connected.

Finally, we also used a technique to investigate mean changes in the ERI-Q constructs, while controlling for changes in factor variances, factor covariances, and factor loadings over time. First of all, traditional repeated-measures MANOVAs revealed mean changes of efforts, esteem rewards, financial/career-related aspects, job security, and overcommitment over time (i.e., alpha change). However, the pattern of change for each construct was opposite in both samples. Whereas efforts increased and the three rewards constructs decreased in Sample 1 (negative changes in general), they successively decreased and increased in Sample 2 (positive changes in general). Moreover, overcommitment decreased over time only in Sample 2. As mentioned before, Sample 1 employees reported frictions due to organizational renewal and personnel changes between the two waves, which could be an explanation for the negative alpha changes. Contrarily, Sample 2 management was willing to improve the job conditions of their workers as a result of the first-wave survey findings. Furthermore, these MANOVA patterns of mean changes were also present when latent means using LISREL were studied (but then controlling for the influence of gamma- and beta-change presence). Another explanation for these findings could be that the difference in time lag between the two samples might have caused the differences. Simulation studies showed that shorter time lags (i.e., one year or less) differ substantially from longer time lags (i.e., two years or more; cf., Zapf, Dormann, & Frese, 1996). For instance, it could be that employees in the two-year study had better opportunities to adapt to their working conditions, which resulted in positive changes. Notwithstanding, the present findings suggest that moderate changes in the ERI-Q factor structure did *not* affect the interpretation of changes in mean efforts, mean rewards, and mean overcommitment in both samples. Therefore, at the measurement level, we can confidently argue that the ERI-Q scales constitute construct-valid measures. This is very important for people who intend to conduct intervention studies using the ERI-Q scales.

Although future research is obviously needed, the present findings extend the existing ERI-Q research by providing new information about issues inherent in job stress research. First, we conclude that the ERI-Q is a psychometrically well-justified measure of assessing psychosocial stress at work. Our demonstration of time invariance of the ERI-Q scales allows researchers to conduct job stress intervention studies without having to worry about whether observed differences and real changes in efforts, rewards, and overcommitment are due to structural change of the ERI-Q scales over time. However, we encourage researchers to examine the assumption of stable instruments before conducting substantive analyses of intervention changes. Although there is still room to sophisticate the ERI-Q, as described earlier (i.e., item similarities of both efforts and overcommitment, the stability of the reward constructs, as well as the way of response rating), the one-step response format in the latest version seems to be applicable to all kind of working populations. Finally, the methodological approaches proposed here offer several benefits in studying the stability of ERI constructs and operationalizations over time. Moreover, it extends it into the job stress and health area where this kind of conceptual advance has not been widely tested.

References

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. Berlin: Springer-Verlag.

Dormann, C., & Zapf, D. (2002). Social stressors at work, irritation, and depressive symptoms: Accounting for unmeasured third variables in a multi-wave study. *Journal of Occupational and Organizational Psychology, 75*, 33–58.

Dragano, N., von dem Knesebeck, O., Rödel, A., & Siegrist, J. (2003). Psychosoziale Arbeitsbelastungen und muskuloskeletale Beschwerden: Bedeutung für die Prävention [Psychosocial work stress and musculoskeletal diseases: Significance for prevention]. *Zeitschrift für Gesundheitswissenschaften, 11*, 196–207.

Frese, M., & Zapf, D. (1988). Methodological issues in the study of work stress: Objective vs subjective measurement of work stress and the question of longitudinal studies. In C. L. Cooper & R. Payne (Eds.), *Causes, coping and consequences of stress at work* (pp. 375–411). Chichester, UK: Wiley & Sons.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12*, 133–157.

Goodman, J., & Blum, T. (1996). Assessing the non-random sampling effects of subject attrition in longitudinal research. *Journal of Management, 22*, 627–652.

Hair, Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis (Sixth Edition)*. Upper Saddle River, NJ: Pearson Education.

Hanson, E. K. S., Schaufeli, W., Vrijkotte, T., Plomp, N. H., & Godaert, G. L. R. (2000). The validity and reliability of the Dutch effort-reward imbalance questionnaire. *Journal of Occupational Health Psychology, 5*, 142–155.

Joksimovic, L., Starke, D., von dem Knesebeck, O., & Siegrist, J. (2002). Perceived work stress, overcommitment, and self-reported musculoskeletal pain: A cross-sectional investigation. *International Journal of Behavioral Medicine, 9*, 122–138.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide (2nd Edition)*. Chicago: Scientific Software International.

Marmot, M., Siegrist, J., & Theorell, T. (2006). Health and the psychosocial environment at work. In M. Marmot & R. G. Wilkinson (Eds.), *Social determinants of health (2nd Edition)* (pp. 97–130). Oxford: Oxford University Press.

Peter, R., Alfredsson, L., Hammar, N., Siegrist, J., Theorell, T., & Westerholm, P. (1998). High effort, low reward, and cardiovascular risk factors in employed Swedish men and women: Baseline results from the WOLF study. *Journal of Epidemiology and Community Health, 52*, 540–547.

Rödel, A., Siegrist, J., Hessel, A., & Brähler, E. (2004). Fragebogen zur Messung beruflicher Gratifikationskrisen [Questionnaire for measuring occupational reward problems]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 25*, 227–238.

Schaubroeck, J., & Green, S. G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. *Journal of Applied Psychology, 74*, 892–900.

Siegrist, J. (1996). Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology, 1*, 27–41.

Siegrist, J., & Peter, R. (1996). Measuring effort-reward imbalance at work: Guidelines. Düsseldorf: Heinrich Heine University.

Siegrist, J., Siegrist, K., & Weber, I. (1986). Sociological concepts in the etiology of chronic disease: The case of ischemic heart disease. *Social Science & Medicine, 22*, 247–253.

Siegrist, J., Starke, D., Chandola, T., Godin, I., Marmot, M., Niedhammer, I., & Peter, R. (2004). The measurement of effort-reward imbalance at work: European comparisons. *Social Science & Medicine, 58*, 1483–1499.

Tsutsumi, A. (2004). *Development and application of the Japanese version of effort-reward imbalance questionnaire*. Okayama, Japan: Okayama University School of Medicine and Dentistry.

Tsutsumi, A., Ishitake, T., Peter, R., Siegrist, J., & Matoba, T. (2001). The Japanese version of the Effort-Reward Imbalance Questionnaire: A study in dental technicians. *Work and Stress, 15*, 86–96.

Tsutsumi, A., & Kawakami, N. (2004). A review of empirical studies on the model of effort-reward imbalance at work: Reducing occupational stress by implementing a new theory. *Social Science & Medicine, 59*, 2335–2359.

Tsutsumi, A., Nagami, M., Morimoto, K., & Matoba, T. (2002). Responsiveness of measures in the effort-reward imbalance questionnaire to organizational changes: A validation study. *Journal of Psychosomatic Research, 52*(4), 249–256.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitments to the organization during the first 6 months of work. *Journal of Applied Psychology, 78*, 557–568.

van Vegchel, N., de Jonge, J., Bosma, H., & Schaufeli, W. (2005). Reviewing the effort-reward imbalance model: Drawing up the balance of 45 empirical studies. *Social Science & Medicine, 60*, 1117–1131.

van Vegchel, N., de Jonge, J., Bakker, A. B., & Schaufeli, W. B. (2002). Testing global and specific indicators of rewards in the Effort-Reward Imbalance Model: Does it make any difference? *European Journal of Work and Organizational Psychology, 11*, 403–421.

Zapf, D., Dormann, C., & Frese, M. (1996). Longitudinal studies in organizational stress research: A review of the literature with reference to methodological issues. *Journal of Occupational Health Psychology, 1*, 145–169.