Scandinavian Journal of Public Health, 2010; 38(Suppl 3): 25–32

ORIGINAL ARTICLE

Reliability of the Copenhagen Psychosocial Questionnaire

SANNIE VESTER THORSEN & JAKOB BUE BJORNER

National Research Centre for the Working Environment, Copenhagen, Denmark

Abstract

Aims: Reliabilities of the work environment questionnaire Copenhagen Psychosocial Questionnaire (COPSOQ) have previously been estimated by Cronbach's alpha, but since the internal consistency assumption may not apply to all COPSOQ scales, Cronbach's alpha may underestimate true reliability. This study aims to evaluate reliability in a test–retest design. *Methods:* We analyzed postal questionnaire data from 349 persons (of whom 283 were employees) who completed two forms with a median interval of 22 (range 6–65) days between baseline and follow-up. Test–retest reliabilities were estimated by the intraclass correlation (ICC). For scales where the internal consistency assumption was theoretically plausible, reliabilities were also estimated by Cronbach's alpha and by Green's test–retest alpha. *Results:* With one exception, the ICC estimated reliabilities of the COPSOQ scales were adequate or good (range 0.70–0.89). A scale concerning mutual trust between employees had a low reliability of 0.64. Among the scales where the internal consistency assumption was plausible, Cronbach's alpha was adequate or good (0.75–0.85) for seven out of eight scales. Green's retest alpha was adequate or good for six out of eight scales (0.72–0.81). ***Conclusions:*** **Standard criteria for acceptable intraclass correlation reliability were achieved for all COPSOQ scales but one. The test–retest design and intraclass correlation appears to be more appropriate than Cronbach's alpha for assessing the reliability of psychosocial work environment scales.**

Key Words: *Copenhagen Psychosocial Questionnaire, Cronbach's alpha, health psychology, occupational, questionnaires, reliability (epidemiology), work environment*

Background

The reliability of a questionnaire scale reflects the amount of variance in the scale that is explained by the construct that scale is intended to measure as opposed to random error (page 28 in [1]]. If a scale that is valid but unreliable is used as an endpoint in statistical analyses, the statistical power will be low [2]. If such a scale is used as an independent variable, results will be biased [3]. Therefore, reliability is an important measurement property of any scale used in research. For assessment of individuals, even higher precision is required than necessary for the group level analyses typically used in research [4].

The most frequently used reliability estimator is Cronbach's alpha [5]. This estimator is easy to achieve since it only requires cross-sectional data. However, all reliability estimators are calculated under a number of assumptions, and the assumptions of a chosen estimator should be carefully evaluated in the particular situation it is used. Cronbach's alpha assumes "internal consistency" [6] and that item specific errors are uncorrelated [7]. These assumptions are crucial to the use of Cronbach's alpha [6,8–11]. A third assumption is that items in a scale have the same true score (tau equivalency) [1]. Due to violation of this less crucial assumption, alpha is sometimes considered a lower bound of reliability.

The "internal consistency assumption" is the assumption that items belonging to the same scale have to have a positive correlation because they are effect indicators of a common unidimensional cause [6,8]. As discussed by Bjorner et al in this issue [9], this assumption seems reasonable for some COPSOQ constructs (e.g. depression, burnout and stress) but not for other constructs (e.g. possibilities for development, recognition, social inclusiveness). For such scales, where items are combined because of

Correspondence: Sannie V. Thorsen, National Research Centre for the Working Environment, Lersø Parkallé 105, DK2100 Copenhagen, Denmark. Tel: +45 3916 5200. Fax: +45 3916 5201. E-mail: svt@arbejdsmiljoforskning.dk

(Accepted 31 August 2009)

DOI: 10.1177/1403494809349859

their hypothesized common effect rather than their common cause, high inter-item correlation is not a necessary criterion of construct validity [6,8]. For such scales with "causal indicators", Cronbach's alpha might not be a good measure of reliability [8], because it might underestimate true reliability. Cronbach's alpha is sometimes interpreted as a measure of "internal consistency", i.e. that the items measure one unidimensional construct. However, this use of Cronbach's alpha is also doubtful since simulation studies show that Cronbach's alpha is not a good indicator of unidimensionality [12].

Another potential problem of Cronbach's alpha is the assumption that item specific errors are uncorrelated, i.e. that all correlation between items is due only to the latent construct measure by the scale. Such violation may be caused by local item dependence [13] (caused, for example, by similarities in wording between two items) or by transient errors (i.e. errors due to specific conditions at the time the test is taken, e.g. the mood of the test-person at the particular time). In both cases, Cronbach's alpha may overestimate true reliability.

The aim of the current study is to estimate reliability of the Copenhagen Psychosocial Questionnaire version II (COPSOQ II) scales, using a test–retest design to avoid the problems of Cronbach's alpha. For the scales that fulfil the internal consistency assumption, we report three reliability estimates; the intraclass coefficient (ICC), Cronbach's alpha and Green's test–retest alpha. Although ICC reliabilities avoid some of the assumptions of Cronbach's alpha, they have assumptions of their own – in particular that the 1) concept measured is stable in the interval between test and retest, and 2) that the error terms are not correlated over time. If the concept measures changes over time, the test–retest design might underestimate the reliability and a cross-sectional assessment of reliability (e.g. Cronbach's alpha) may be preferable. If the assumption of uncorrelated error terms between baseline and follow-up is not fulfilled, the test–retest design might overestimate the reliability. An example of correlated error terms in test–retest design is when the respondents' answers at time two are influenced by their recollection of their responses to the same items at time one, the "carry-over effect" [14]. Finally, the answers in the second test may also be influenced by "practice effect", i.e. that the respondent learns how to answer the questionnaire the first time [14]. For scales where the internal consistency assumption applies, Green's retest alpha is an alternative to the ICC and Cronbach's alpha reliability estimators. The Green's test–retest alpha estimator is similar to Cronbach's alpha but removes transient error from the estimate by using two time points in the calculation [10]. It also takes problems due to "carry-over effect" into account because it compares different items at the two time points. For further comparisons, we report Green's reliabilities for the scales where we find the internal consistency assumption to be plausible.

Methods

The study was designed as a test–retest study using postal questionnaires. A random sample of 1,000 Danes between 18-59 years, selected from the CPR registry, received the baseline questionnaire and an introductory letter explaining that respondents would receive another questionnaire with similar questions shortly after answering the first questionnaire. Two weeks after receiving the first questionnaire, we mailed a second questionnaire to all responders. A letter included with the retest questionnaire instructed the respondent to answer not from memory of the answers to the first test but to choose the answer that best described his or her current situation. Non-responders to the first or the second questionnaire received two reminders by post, unless they had indicated that they did not want to take part in the study. The study aimed to achieve test and retest responses from at least 200 employees, which would result in a confidence interval of 0.68–0.81 for a scale with a reliability of 0.75. The study exceeded these minimum requirements, achieving responses from 457 persons for the baseline questionnaire and from 349 persons for the follow-up questionnaire, 283 of whom were currently employed (see Table I).

Questionnaire

The study used the medium length version of the COPSOQ II (Pejtersen et al in this issue [18]) to avoid excessive response burden. Abbreviated content of the items can be seen in Table II. The total number of items was 112 in the first questionnaire and 97 in the second questionnaire. The questionnaire included the standard 23 scales of the medium length COPSOQ plus two extra scales from the full length questionnaire. We added the scales on *Demands for hiding emotions* and *Social inclusiveness* based on the hypothesis that they did not fulfil the internal consistency assumption and thus that previous reliability estimates might be biased. Also, based on the same logic we included the question "*Is your salary fair in relation to your effort at work?*" that was

Table I. Sample characteristics.

	First questionnaire ($n = 457$)	Both questionnaires ($n = 349$)	Employees[a] ($n = 283$)
Age of respondents Median (range)	42 (18–58)	43 (18–58)	44 (19–58)
Days between assessments. Median (range)	–	22 (6–65)	22 (6–65)
Percent males	42%	41%	42%
Percent blue collar[b]	34%[b]	32%[b]	32%[b]

[a]Not including self employed. [b]Percentage of blue collar workers to blue and white collar. Unemployed and self employed are not included. Data for blue/white collar status is available for 330 employees who answered the first questionnaire and for 258 employees who answered both questionnaires.

part of the "Recognition" scale in the COPSOQ II test version but was subsequently removed due to low correlations with other items. The internal consistency assumption was found to be reasonable for eight scales (see Table II). The assessment was based on considerations regarding the items' status as causes or effects of the latent construct. The construct lacks internal consistency if at least one item is considered a cause [6,8].

Statistical analysis

All COPSOQ scales were calculated using the standard recommendations of simple sum scoring and transformation to a 0–100 metric (see Pejtersen et al in this issue [18]). An average value of each scale's score is given in Table III. Test–retest reliabilities were calculated using an intraclass correlation [15] for all 25 scales. We identified eight scales where the internal consistency seemed plausible (Table II). For these scales Cronbach's alpha and Green's retest alpha were calculated. To enable comparisons, all reliability coefficients were calculated only for those respondents who had completed the questionnaire both at time one and two. Cronbach's alpha was calculated separately from baseline data and from retest data. We used empirical bootstrap [16] to estimate confidence intervals for all reliability estimates (1,000 bootstrap samples) and to test whether reliabilities were significantly different across age, gender, white/blue collar, and time between assessments. For these comparisons we dichotomized age and time between assessments at their median value.

We used SAS version 9.1 for all analyses and wrote an SAS macro to calculate Green's test–retest coefficient alpha (available from the authors).

Results

The respondent rate was 46% for the first questionnaire and 35% for the second questionnaire. Of these, 283 (81%) were employees at a workplace (as opposed to being unemployed or self-employed). Sample characteristics are shown in Table I. The time interval between test and retest responses of the questionnaire was approximately three weeks. The range was from 6 days to 65 days. Time intervals shorter than 14 days were possible because a few respondents answered the first questionnaire late and then answered a questionnaire (actually intended to be the first) sent out with a reminder.

Intraclass correlation provided acceptable to high reliability estimates for 24 out of 25 scales (ICC range 0.70–0.89, see Table III). The scale concerning mutual trust between employees had an ICC reliability of 0.64. For the eight scales where we found the assumption of internal consistency reasonable, Cronbach's alpha was generally high, although one scale (*Meaning of work*) had a reliability of 0.68. Among these eight scales, ICC reliability estimates were higher than Cronbach's alpha for the scales concerning the psychosocial working environment or the work-individual interface (*Meaning of work*, *Commitment to the workplace, Role clarity,* and *Work–family conflict*), while Cronbach's alpha was higher for scales measuring health outcomes (*Burnout, Stress* and *Sleep troubles*). Green's retest alpha was lowest for all scales.

On average, the transient "error" (the difference between Cronbach's alpha and Green's retest alpha) was 0.07 (range 0.02–0.18). The three scales with the highest transient dependency were the scales regarding health (*Stress*, *Burnout* and *Sleep problems*), while the scales concerning the working environment had lower transient "error".

Cronbach's alpha reliabilities of all eight scales increased from test data to retest data (results not shown). The average increase was 0.03 (range 0.02–0.06).

The ICC reliabilities calculated from respondents with short time intervals between responses was compared to the reliability from respondents with larger time intervals between responses.

Table II. Internal consistency of scales.

Scale (number of items)	Consistency	Items in scale: *causal indicators* and effect indicators[a]
Demands at work		
Quantitative demands (4)	No	QD1 *Work piles up*; QD2 Complete task; QD3 Get behind; QD4 Enough time
Work pace (3)	Yes	WP1 Work fast; WP2 High pace; WP3 High pace necessary
Emotional demands (4)	No	ED1 Emotional disturbing; ED2 *Relate to other people's problems*; ED3 Emotionally demanding; ED4 Emotionally involved
Demands for hiding emotions (3)	No	HE1 *Treat equally*; HE2 Hide feelings; HE 3 *Kind and open*
Work organization and job contents		
Influence (4)	No	IN1 Influence work; IN2 *Say in choosing colleges*; IN3 *Amount of work*; IN4 *Influence work task*
Possibilities for development (4)	No	PD1 *Take initiative*; PD2 Learning new things; PD3 *Use skills*; PD4 Develop skills
Meaning of work (3)	Yes	MW1 Work meaningful; MW2 Work important; MW3 Motivated and involved
Commitment to the workplace (4)	Yes	CW1 Enjoy telling others; CW2 Workplace great importance; CW3 Recommend a friend; CW4 Looking for work elsewhere
Interpersonal relations and leadership		
Predictability (2)	No	PR1 *Informed about changes*; PR2 *Information to work well*
Recognition (rewards) (4)	No	RE1 *Recognised by management*; RE2 *Respected by management*; RE4 *Fair salary*; RE5 *Respected by colleagues*[b]
Role clarity (3)	Yes	CL1 Clear objectives; CL2 Responsibility; CL3 Expectation
Role conflicts (4)	No	CO1 *Mixed acceptance*; CO2 *Contradictory demands*; CO3 *Do things wrongly*; CO4 *Unnecessary tasks*
Quality of leadership (4)	No	QL1 *Development opportunities*; QL2 *Prioritise job satisfaction*; QL3 *Work planning;* QL4 *Solving conflicts*
Social support from supervisor (3)	No	SS2 Support supervisor; SS1 *Supervisor listens to problems*; SS3 *Supervisor talks about performance*
Social support from colleagues (3)	No	SC1 Support colleagues; SC2 *Colleagues listen to problems*; SC3 *Colleagues talk about performance*
Social community at work (3)	No	SW1 *Atmosphere*; SW2 *Cooperation*; SW3 Community
Work–individual interface		
Job satisfaction (4)	No	JS1 *Work prospects*; JS2 *Work conditions*; JS3 *Work abilities*; JS4 Job in general
Work–family conflict (4)	Yes	WF1 Being both places; WF2 Energy conflict; WF3 Time conflict; WF4 Family think you work too much
Values at the workplace		
Trust regarding management (4)	No	TM1 *Management trust employees*; TM2 *Employees trust information*; TM3 *Management withhold information*; TM4 *Employees express views*
Mutual trust between employees (3)	No	TE1 *Colleagues withhold information*; TE2 *Withhold information management*; TE3 Trust colleagues
Justice (4)	No	JU1 *Conflicts resolved fairly*; JU2 *Employees appreciated*; JU3 *Suggestions treated seriously*; JU4 *Work distributed fairly*
Social inclusiveness (3)	No	SI1 *Gender discrimination*; SI2 *Race/religion discrimination*; SI3 *Age discrimination*; SI4 *Health discrimination*
Health and well-being		
Self-rated health (1)	No	GH1 Health all in all
Burnout (4)	Yes	BO1 Worn out; BO2 Physically exhausted; BO3 Emotionally exhausted; BO4 Tired
Stress (4)	Yes	ST1 Problems relaxing; ST2 Irritable; ST3 Tense; ST4 Stressed
Sleeping troubles (4)	Yes	SL1 Slept badly; SL2 Hard to sleep; SL3 Woken up early; SL4 Woken up several times

[a]Each item/question in the scale is given in abbreviated form. Assumed causal items are written in italics. [b]A non-standard version of the scale was evaluated.

Only one difference had a p value below 0.05 and it was insignificant after Bonferroni adjustment (see Table IV).

Similar tests were performed for age and for blue/white collar workers and for gender (see Table IV). We found four tests to be below p value 0.05 and one of the p values could pass a Bonferroni adjustment test. The scale *Burnout* had significant higher reliability for older respondents compared to younger ($p = 0.0006$).

Discussion

The design of the current study was prompted by the hypothesis that the internal consistency approach to estimation of reliability of the COPSOQ scales may

Table III. Reliability estimates for COPSOQ scales.

Scale	*n*	Internal consistency	Score time-one (*std*)	Score time-two (*std*)	ICC (*95% CI*)	Cronbach's alpha (*95% CI*)	Green's retest alpha (*95% CI*)
Demands at work							
Quantitative demands	271	No	40 (*20*)	38 (*19*)	0.87 (*0.83–0.89*)		
Work pace	275	Yes	54 (*20*)	53 (*20*)	0.85 (*0.81–0.88*)	0.85 (*0.82–0.88*)	0.81 (*0.77–0.85*)
Emotional demands	275	No	38 (*23*)	38 (*24*)	0.89 (*0.86–0.91*)		
Demands for hiding emotions	268	No	50 (*21*)	53 (*22*)	0.75 (*0.70–0.80*)		
Work organization and job contents							
Influence	270	No	50 (*22*)	51 (*21*)	0.83 (*0.78–0.87*)		
Possibilities for development	272	No	33 (*17*)	34 (*17*)	0.80 (*0.76–0.84*)		
Meaning of work	275	Yes	25 (*14*)	27 (*14*)	0.74 (*0.67–0.79*)	0.68 (*0.59–0.74*)	0.61 (*0.52–0.68*)
Commitment to the workplace	271	Yes	39 (*19*)	39 (*19*)	0.87 (*0.83–0.90*)	0.75 (*0.67–0.80*)	0.72 (*0.66–0.78*)
Interpersonal relations and leadership							
Predictability	274	No	44 (*21*)	45 (*20*)	0.70 (*0.63–0.77*)		
Recognition (reward)	274	No	36 (*16*)	36 (*16*)	0.80 (*0.76–0.84*)		
Role clarity	271	Yes	29 (*16*)	31 (*17*)	0.80 (*0.76–0.84*)	0.77 (*0.71–0.82*)	0.72 (*0.65–0.77*)
Role conflicts	266	No	36 (*17*)	38 (*18*)	0.74 (*0.68–0.80*)		
Quality of leadership	222	No	43 (*20*)	43 (*20*)	0.83 (*0.77–0.87*)		
Social support from supervisor	226	No	38 (*22*)	37 (*22*)	0.73 (*0.64–0.80*)		
Social support from colleagues	256	No	40 (*19*)	42 (*19*)	0.70 (*0.62–0.76*)		
Social community at work	268	No	20 (*17*)	24 (*17*)	0.73 (*0.66–0.79*)		
Work–individual interface							
Job satisfaction	247	No	25 (*13*)	26 (*12*)	0.73 (*0.64–0.80*)		
Work-family conflict	273	Yes	24 (*17*)	24 (*18*)	0.86 (*0.82–0.89*)	0.80 (*0.75–0.84*)	0.76 (*0.70–0.81*)
Values at the workplace							
Trust regarding management	263	No	32 (*16*)	33 (*16*)	0.80 (*0.74–0.84*)		
Mutual trust between employees	265	No	29 (*16*)	30 (*16*)	0.64 (*0.55–0.73*)		
Justice	268	No	41 (*16*)	40 (*16*)	0.79 (*0.73–0.84*)		
Social inclusiveness	260	No	28 (*15*)	29 (*15*)	0.75 (*0.69–0.80*)		
Health and well-being							
Self-rated health	347	No[a]	40 (*23*)	41 (*21*)	0.78 (*0.72–0.82*)		
Burnout	344	Yes	22 (*19*)	23 (*19*)	0.79 (*0.75–0.83*)	0.81 (*0.77–0.84*)	0.72 (*0.66–0.76*)
Stress	345	Yes	28 (*18*)	28 (*18*)	0.72 (*0.65–0.78*)	0.85 (*0.81–0.88*)	0.67 (*0.60–0.73*)
Sleeping troubles	343	Yes	22 (*19*)	23 (*19*)	0.81 (*0.76–0.85*)	0.84 (*0.80–0.87*)	0.74 (*0.68–0.79*)

[a]Self-rated health only includes one item.

underestimate true reliability, because many of the indicators of the psychosocial working environment must be considered causal indicators rather than effect indicators [6,8]. Previous studies of the COPSOQ using Cronbach's alpha (Pejtersen et al in this issue [18]) have found adequate reliability for most scales, but low alphas for the following scales from the medium length COPSOQ: *Demands for hiding emotions* ($p = 0.57$), *Role conflicts* ($\rho = 0.67$), and *Social inclusiveness* ($\rho = 0.63$). We found higher ICC retest reliabilities (0.75, 0.74, and 0.75, respectively) for these scales, which could suggest that Cronbach's alpha coefficient provided biased estimates of reliabilities for these scales.

The average ICC test–retest reliability across all the "not internally consistent" scales was only slightly higher than the average Cronbach's alpha coefficient previously reported for the same scales (and also only slightly higher than Cronbach's alpha for the "not internally consistent" scales when estimated in this study – data not shown). This may appear to be in contrast with our argument about the dangers of using Cronbach's alpha on "not internally consistent" scales. However, since item selection for the COPSOQ scales was partly determined by the ability of the items to increase alpha (Pejtersen et al in this issue [18]), Cronbach's alpha may be high for the COPSOQ scales, even in cases where the internal consistency assumption is not theoretically plausible. There is nothing inherently wrong in choosing items to maximize Cronbach's alpha in a scale without internal consistency. However, such an approach may put an unnecessary constraint on the scale by rejecting good causal items because they fail to fulfil an unnecessary constraint. In turn, this may hamper content validity and predictive validity.

Although lack of internal consistency may cause Cronbach's alpha to be lower than the true reliability, transient error and local dependence may inflate alpha. Green's retest alpha removes the transient errors from the estimate, but is similar to Cronbach's alpha in other respects. However, both Green's retest

Table IV. Reliability (ICC) estimates for subgroups differing in response time-interval, age, blue/white collar or gender.

	Interval time		Age				Gender	
Scale	6–22 days	23–65 days	18–42 years	43–58 years	Blue collar	White collar	Male	Female
Demands at work								
Quantitative demands	0.88	0.85	0.87	0.86	0.81	0.88	0.86	0.88
Work pace	0.88	0.83	0.85	0.85	0.86	0.82	0.87	0.84
Emotional demands	0.90	0.88	0.90	0.89	0.88	0.88	0.87	0.89
Demands for hiding emotions	0.75	0.76	0.78	0.74	0.84*	0.68*	0.75	0.76
Work organization and job contents								
Influence	0.82	0.84	0.79	0.85	0.82	0.83	0.86	0.80
Possibilities for development	0.84*	0.75*	0.81	0.80	0.78	0.78	0.79	0.81
Meaning of work	0.77	0.71	0.79	0.69	0.72	0.76	0.74	0.73
Commitment to the workplace	0.88	0.84	0.85	0.88	0.87	0.87	0.86	0.87
Interpersonal relations and leadership								
Predictability	0.66	0.75	0.71	0.70	0.70	0.69	0.72	0.69
Recognition (rewards)	0.84	0.77	0.79	0.82	0.86	0.79	0.81	0.80
Role clarity	0.81	0.80	0.81	0.79	0.73	0.83	0.81	0.80
Role conflicts	0.81	0.72	0.74	0.75	0.72	0.77	0.72	0.74
Quality of leadership	0.81	0.86	0.83	0.82	0.86	0.82	0.80	0.83
Social support from supervisor	0.75	0.75	0.72	0.73	0.77	0.70	0.70	0.75
Social support from colleagues	0.74	0.65	0.66	0.72	0.76	0.66	0.66	0.71
Social community at work	0.75	0.72	0.77	0.71	0.67	0.77	0.68	0.78
Work–individual interface								
Job satisfaction	0.73	0.72	0.78	0.71	0.70	0.77	0.77	0.70
Work-family conflict	0.87	0.82	0.86	0.87	0.78*	0.88*	0.87	0.85
Values at the workplace								
Trust regarding management	0.83	0.77	0.73	0.84	0.85	0.77	0.84	0.76
Mutual trust between employees	0.53	0.72	0.61	0.67	0.63	0.66	0.60	0.67
Justice	0.77	0.81	0.80	0.79	0.87	0.79	0.84	0.76
Social inclusiveness	0.74	0.76	0.73	0.76	0.77	0.74	0.79	0.72
Health and well-being								
Self-rated health	0.78	0.77	0.70	0.81	0.76	0.75	0.75	0.79
Burnout	0.82	0.76	0.70*	0.85*	0.77	0.78	0.78	0.80
Stress	0.73	0.72	0.74	0.71	0.62	0.77	0.73	0.72
Sleeping troubles	0.84	0.79	0.71*	0.85*	0.79	0.82	0.78	0.82

*Pairs of reliability estimates with an asterisk are significantly different ($p < 0.05$, no Bonferroni adjustment).

alpha and the ICC test–retest reliability coefficient assume that there is no true change in the domains measured. For many COPSOQ domains, such assumption is reasonable within the retest lag used in this study, but change may well occur within some domains, e.g. stress and sleeping troubles. We suspect this is the reason that test–retest reliability is lower than Cronbach's alpha for the three health-related domains (i.e. the "transient error" is a true change and a one time point assessment is the optimal). In our opinion, Green's retest alpha represents a lower bound estimate of reliability, since the coefficient is unaffected by most of the factors that may cause upwards bias in reliability estimates, but still prone to factors that may cause downwards bias. Thus, it is unsurprising that the retest alpha provides the lowest reliability estimate for all scales where it is used.

The test–retest design may introduce errors due to a "practice effect" or a "carry-over effect". In this context, a practice effect means that respondents achieve a better understanding of the items by answering them in round one and that this makes their answers in round two more reliable. This would again lead to increased Cronbach's alpha reliability in round two and to some (although not as large) increase in the ICC and Green's retest alpha reliability. We did find slightly higher Cronbach's alpha values at time two compared to time one. The differences were fairly consistent and on average 0.03. This would lead to a small (app. 0.015) upward bias in the ICC test–retest coefficient as a measure of scale reliability for a person who has never answered the questionnaire before.

A carry-over effect means that a person remembers his or her last answers and copies these answers in the retest. The round two introduction letter instructed the respondent not to take time one answers into account, but nevertheless a carry-over effect cannot be ruled out. A carry-over effect may influence the

ICC but not Green's retest alpha, because covariances in this estimate are between different times *and* different items. For the five work environment scales where Green's retest alpha was calculated, we found that Green's retest alpha was 0.04 to 0.13 lower than the ICC estimate. This might be due to a carry-over effect or occur because our assumption of internal consistency and unidimensionality does not hold. The standard criterion for reliability is 0.70. Green's retest alpha exceeded 0.70 for four of the five work environment scales, but not for the scale *Meaning of work* (Green's retest alpha = 0.61). The ICC is 0.74 for this scale, while Cronbach's alpha was 0.68 at time one and 0.74 at time two. A previous study (Pejdersen et al in this issue) found a Cronbach's alpha of 0.74. The low Green's retest alpha for the scale *Meaning of work* could be a true low reliability of the scale, but since Cronbach's alpha is lower at baseline than in other studies and data sets, it is likely that the low retest alpha is an aberrant result. Another explanation is that the internal consistency/unidimensionality assumption does not hold for the scale, in which case the ICC retest reliability estimate is the most appropriate estimate for this scale.

While a reliability of 0.70 is traditionally regarded a threshold for adequate reliability for group analyses [4], we would like to emphasize that this threshold is arbitrary and that adequate reliability depends on the study and the purpose. For scales that are used as endpoint variables, low reliability will not bias the estimates, but merely weaken the power of the study [2], an effect that can be countered by increasing the sample size. For scales that are used as independent variables, low reliabilities may bias results, but this effect may be countered by applying analytic methods, e.g. structural equation models, that can take measurement error into account [17]. Thus, while high reliability is preferable, sensible statistical analyses can still be made with scales that have less than perfect reliability. Even higher reliability is required for assessment of individual respondents, but COPSOQ has rarely been used for that purpose.

Conclusion

Using a test–retest design, we found ICC test–retest reliabilities that according to standard guidelines were adequate to good in 24 out of 25 scales in the medium length COPSOQ questionnaire. In particular, we found adequate reliabilities for scales that had low reliability in previous studies based on Cronbach's alpha. For eight scales considered to be internally consistent Cronbach's alpha and Green's retest alpha were calculated and one scale is below standard recommendations for both "internal consistency" estimates. We found indications of practice effects and transient differences between scale estimates of test and retest. It appears that older respondents have higher ICC reliability than younger respondents of the scale *Burnout*. Based on conceptual considerations and the results of the current study, we recommend the ICC coefficient as a reliability estimate and that the internal consistency assumptions are carefully considered before Cronbach's alpha is used to estimate reliabilities for scales concerning the psychosocial working environment.

Acknowledgements

This study was conducted under funding from the Danish Ministry of Employment. The authors wish to thank Pia Gøtterup, Christian Trolle Strandfelt, Christian Roepstorff, and Helle Soll-Johanning for help with data collection and thank Reiner Rugulies and two anonymous reviewers for helpful comments on a previous version of the paper.

References

[1] Thissen D, Wainer H. True score theory: the traditional method. In: Thissen D, Wainer H, editors. Test scoring, Lawrence Erlbaum Associates, Publishers; 2001. pp. 23–71.
[2] Kraemer HC. To increase power in randomized clinical trials without increasing sample size. Psychopharmacol Bull 1991;27(3):217–24.
[3] Carroll RJ. Covariance analysis in generalized linear measurement error models. Stat in Med 1989;8(9):1075–93.
[4] Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw-Hill, Inc.; 1994.
[5] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16(3):297–334.
[6] Bollen KA. Multiple indicators – internal consistency or no necessary relationship. Quality & Quantity 1984;18(4):377–85.
[7] Green SB, Lissitz RW, Mulaik SA. Limitations of coefficient alpha as an index of test unidimensionality. Educ Psychol Measure 1977;37(4):827–38 .
[8] Bollen K, Lennox R. Conventional wisdom on measurement – a structural equation perspective. Psychol Bull 1991;110(2):305–14.
[9] Bjorner JB, Pejtersen JH. Evaluating construct validity of the Copenhagen Psychosocial Questionnaire through analysis of differential item functioning and differential item effect. Scand J Public Health 2010;38(Suppl 3):90–105.
[10] Green SB. A coefficient alpha for test-retest data. Psychol Meth 2003;8(1):88–101.
[11] Osburn HG. Coefficient alpha and related internal consistency reliability coefficients. Psychol Meth 2000;5(3):343–55.
[12] Hattie JA. Methodology review: assessing unidimensionality of tests and items. Appl Psychol Measure 1985;9(2):139–64.
[13] Chen W-H, Thissen D. Local dependence indexes for item pairs using item response theory. Educ Behav Stat 1997;22(3):265–89.

[14] Mckelvie SJ. Does memory contaminate test-retest reliability. J Gen Psychol 1992;119(1):59–72.
[15] Shrout PE, Fleiss JL. Intraclass correlations – uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8.
[16] Henderson AR. The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. Clin Chim Acta 2005;359(1–2):1–26.
[17] Budtz-Jorgensen E. Estimation of the benchmark dose by structural equation models. Biostatistics 2007;8(4): 675–88.
[18] Pejtersen JH, Kristensen TS, Borg V, Bjorner JB. The second version of the Copenhagen Psychosocial Questionnaire. Scand J Public Health 2010;38(Suppl 3): 8–24.