International Journal of Occupational Safety and Ergonomics (*JOSE*) 2006, Vol. 12, No. 1, 53–68

Reliability of a Questionnaire and an Ergonomic Checklist for Assessing Working Conditions and Health at Call Centres

Kerstin Norman

Department for Work and Health, National Institute for Working Life, Stockholm, Sweden

Håkan Alm

Luleå University of Technology, Luleå, Sweden

Ewa Wigaeus Tornqvist
Allan Toomingas

Department for Work and Health, National Institute for Working Life, Stockholm, Sweden
Department of Public Health Sciences, Karolinska Institute, Sweden

Background. *The purpose was to study the test-retest reliability and internal consistency of questions in a questionnaire concerning working conditions and health and the inter-rater reliability of observations and measurements according to an ergonomic checklist.* ***Method.*** *Fifty-seven operators participated in a retest questionnaire and 58 operators participated in an inter-observer test.* ***Results****. The questions had* fair *to* good *or higher reliability in 142 of the total of 312. Twenty-seven of the total of 44 variables in the ergonomic checklist were classified as having fair to good or higher reliability.* ***Conclusions.*** *About half of the questions had* fair *to* good *or higher reliability and can be recommended for further analyses. The majority of variables in the ergonomic checklist were classified as having* fair *to* good *or higher reliability.* Low *reliability does not necessarily indicate that the reliability of the test, per se, is low but may signify that the conditions measured vary over time or that the answers are aggregated in one part of the scale.*

test-retest reliability inter-observer test questionnaire ergonomic checklist
call centre computer work

1. INTRODUCTION

The foundation of all empirical investigations in ergonomics consists of measurements and observations made on the subjects or objects of interest. Clearly, such measurements need to be objective, precise, and reproducible, for reasons eloquently summarized in the quotation from Fleiss [1].

> The most elegant design of a study will not overcome the damage caused by unreliable or imprecise measurements. The requirement that one's data be of high quality is at least as

Financial support from the Swedish Council for Working Life and Social Research is greatly appreciated. We are also grateful to all call centre companies and subjects who participated in the study, and to everybody in the call centre project group who contributed with highly qualified work.

The work was performed at the Department for Work and Health, National Institute for Working Life, Stockholm, Sweden. Grant sponsor: Swedish Council for Working Life and Social Research. Grant number: FAS Dnr 2001-2812.

Correspondence and requests for offprints should be sent to Kerstin Norman, Department for Work and Health, National Institute for Working Life, Vanadisvägen 9, SE-113 91 Stockholm, Sweden. E-mail: <kerstin.norman@niwl.se>.

> important a component of a proper study design as the requirement for randomisation, double blinding, controlling where necessary for prognostic factors, and so on. Larger sample size than otherwise necessary, biased estimates and even biased samples is some of the untoward consequences of unreliable measurements that can be demonstrated. (p. 2)

In all empirical studies it is therefore important to ensure that the information collected is as accurate as possible. In assessing the accuracy of any particular measuring instrument or method, a common distinction is between the reliability and the validity of the instrument or method [2]. Reliability is essentially the extent of the agreement between repeated measurements of the same material, and validity is the extent to which an instrument or method measures what it is supposed to measure. When it comes to data collection of any kind, for any type of data, qualitative as well as quantitative, it is important to know if the data has the qualities needed for a scientific study. High reliability is needed in order to obtain high validity, but high reliability does not guarantee high validity.

Kerlinger and Lee [3] defined reliability as the accuracy or precision of a measuring instrument. He also suggested some synonyms for reliability: dependability, stability, consistency, predictability, and accuracy. Kerlinger and Lee also discussed the question of reliability using three different approaches. One approach to reliability is to ask the question: If we measure the same set of objects again and again with the same or a comparable measuring instrument, will we get the same or similar results? This question implies a definition of reliability in terms of stability, dependability, and predictability, which is the most common definition in elementary discussions of the subject.

A second approach is to focus on the question if the measures obtained from a measuring instrument are "true" measures of the property measured. This is an accuracy definition. Compared to the first definition, it is further removed from common sense and intuition, but it is also more fundamental. These two approaches or definitions can be summarized in the words stability and accuracy.

A third approach to the definition of reliability is one that not only helps us better define and solve both theoretical and practical problems but also implies other approaches and definitions. We can inquire how systematic or random the error of measurement in a measuring instrument is. Reliability is the accuracy or precision of a measuring instrument.

Some variables in the physical and social environment can be assumed to be rather stable over time and one possibility is to measure the same variables at different times. The correlation between measures taken at different time periods can be used as a measure of reliability. An estimate of this kind is referred to as a coefficient of stability or test-retest reliability. Another measure of reliability uses parallel forms of tests, i.e., tests that cover the same content, use the same type of questions, and are equally difficult. The correlation between the test scores from the different tests is used as a measure of reliability. In this case the reliability question is how consistent these tests are. This method is called parallel forms reliability. A third possibility is to split a test into two equivalent and independent halves and calculate the correlation between these two halves. This method is called a split-half reliability.

Yet another important aspect of reliability is the consistency between different observers, when observing the same phenomenon. This is called inter-rater agreement and can be tested by letting several people observe the same object or process and by comparing their observations.

One common method of calculating the test-retest and inter-observer reliability is to use Pearson's product moment correlation coefficient for association between two continuous variables. Spearman's rank correlation coefficient is used as a measure of the association between two categorical variables, measured at least at an ordinal level. Cohen's Kappa is often used as a reliability coefficient for binary categorical variables. Cohen's Kappa coefficient is a measure of the proportional agreement beyond agreement expected by chance.

Another approach is to focus upon the internal consistency of an instrument and ask the question: Do all items in a test measure the same variable? Or do different groups of items in a measuring instrument actually measure the same variable? For instance, is there agreement in the outcome between different parts of the questionnaire that measure the same phenomena? This is called internal consistency reliability. One common measure of internal consistency is Cronbach's alpha [4].

Call centres are one of the most rapidly growing forms of workplaces in Sweden. According to the state-run Invest in Sweden Agency (ISA), approximately 60,000 people were employed in Swedish call centres in 2002 (www.isa.se[1]). The expansion of call centres can have a positive impact on many communities by creating new jobs. However, problems have been noted, e.g., with the wage, feedback systems (e.g., being monitored by computers that can register performance), working hours, inadequate opportunities for professional development, and insufficient physical and psychosocial conditions and variation in these conditions. There are few studies concerning working conditions and health status at call centres [5, 6] and it is important, at an early stage, to survey the different risks that may occur, so that we can prevent these risks and promote a sustainable working environment.

Now that we have started to study the call centre branch it is important to use instruments of high quality. If the measurements are shown to be reasonably reliable this will increase confidence in the outcome of these studies.

2. AIM

The purpose was to study the test-retest reliability and internal consistency of questions in a questionnaire covering symptoms, physical and psychosocial working conditions at call centres, and also the inter-rater reliability of observations and measurements according to an ergonomic checklist.

3. METHOD

3.1. Overall design

This study represents a cross-sectional study of a selected number of call centre operators in Sweden.

3.2. Companies

A total of 38 call centre companies were invited to participate in the study. The companies were selected to represent different types of call centres: internal and external companies, companies with tasks that varied in the degree of complexity, companies located in large and small cities, different ownership (Swedish public owner, Swedish private owner, and international owner) and different parts of the country. The goal was not to obtain a representative sample of call centres, but to get a basis for comparisons between call centres of different types. Sixteen companies representing 28 different call centre sites agreed to participate in the study.

3.3. Participants

Of the total of 1,802 call centre operators, employed at the 16 companies, 1,531 subjects (984 women and 547 men) fulfilled the inclusion criteria for participation in the investigation (Table 1). The inclusion criteria for participation in the investigation was that the subjects should have worked at the call centre company for at least 1 month, and have had customer contacts. Subjects, who were on sick, holiday, parental, or other leave, as well as those who had quit their employment, were excluded from the study. All included operators were asked to fill in a questionnaire, and 1,183 complete questionnaires were received after two reminder rounds. From the subjects who answered the questionnaire, 71 operators (57 women and 14 men) were randomly selected to participate in a retest round, and 47 women and 10 men responded to the retest questionnaire.

[1] Retrieved January 23, 2004.

TABLE 1. Operators Invited to Answer, and Actual Responders, to the Original and to the Retest Questionnaires. Those Invited, and Actual Participants in Observations According to the Ergonomic Checklist, and in the Inter-Observer Test of the Ergonomic Checklist

	Total		Women		Men	
Invited and Responders/Participants	**(*n*)**	**%**	**(*n*)**	**%**	**(*n*)**	**%**
Invited to answer original questionnaire	1531		984		547	
Actual responders	1183	77	848	86	335	61
Invited to answer retest questionnaire	71		57		14	
Actual responders	57	80	47	82	10	71
Invited to participate in observations according to ergonomic checklist	160		113		47	
Actual participants	159	99	112	99	47	100
Invited to participate in inter-observer test of ergonomic checklist	60		46		14	
Actual participants	58	97	44	96	14	100

Ten operators, from each of the 16 companies who were invited to participate in measurements, were selected by staff managers or coaches. Observations were made by trained ergonomists of working conditions and work postures (Table 1). One operator refused to participate and no substitute could fill in, resulting in a total of 159 operators. The criterion for participation in the ergonomic investigation was that the operator had to work during the 2 days when the ergonomists visited the company. A sub-group of 60 operators (46 women and 14 men) were invited to participate in an inter-observer reliability test. Fifty-eight operators (44 women and 14 men) agreed to participate.

3.4. Material and Instruments

3.4.1. Questionnaire

A questionnaire consisting of 312 questions was used, covering physical and psychosocial working conditions, and musculoskeletal symptoms during the previous month. A complete version of the questionnaire can be found at the web site of the National Institute for Working Life, www.arbetslivsinstitutet.se/datorarbete/pdf/CCBaselineQuest.pdf[2].

The questionnaire comprised questions about background variables, employment, working hours and remuneration (reward), duties, computer work, comfort, disruptions in the computer system and technical support, management, social support and development, job requirements, call logging and monitoring, stress and fatigue, sleep, winding down and recovery, work/life balance, general health, symptoms, measures taken to reduce the symptoms, the way symptoms affected their working capacity, stress-related symptoms, self-reported work-related injuries, absence from work, presence at work despite being ill enough to stay at home, and questions about eye tests and glasses.

Sixteen indices were constructed as arithmetic means of the answers to groups of questions: comfort (a) noise, lighting and air quality (5 questions, No. 38 a–e in the questionnaire) and (b) furniture and equipment (9 questions, No. 38 f–n); social support (7 questions, No. 41 a–g); support from supervisor (8 questions, No. 41 h–o); psychological demands (14 questions, No. 44 a–m, 45 j); cognitive demands (7 questions, No. 44 b, c, f–h, j, l); time pressure (3 questions, No. 44 e, m, 45 j); emotional demands (3 questions, No. 44 a, d, i); lack of control (7 questions, No. 45 c, e, i, l–o); limited decision latitude (4 questions, No. 45 e, m–o); positive work (16 questions, No. 45

[2] Retrieved January 19, 2006.

a–i, k–q); stress (4 questions, No. 48 a, c, f, g); energy (4 questions, No. 48 b, d, e, h); feeling worn out (3 questions, No. 48 i, l, n); anxiousness (4 questions, No. 48 p, r–t); and psychosomatic symptoms (5 questions, No. 62 a–e).

3.4.2. Ergonomic checklist

A checklist was used as a tool to evaluate working conditions in call centre work. A complete version of the ergonomic checklist can be found at the web site of the National Institute for Working Life, www.arbetslivsinstitutet.se/datorarbete/pdf/CCXlist.pdf[3]. The checklist consisted of 14 different parts: dimensions of office, indoor air quality and climate, sound level, electromagnetic fields, illumination, lighting conditions and vision ergonomics, standard of office table and chair, computer equipment and its arrangement, work postures and movements, operator's knowledge about optimal adjustments of furniture and equipment, and working technique.

Inter-observer tests were made of some variables in the ergonomic checklist. Example of variables that were observed was backrest height, if the control device was positioned within forearm's length and shoulder width, the main source of disturbing noise, if there were visible reflections on the desk and working postures. Measurements of luminance and viewing angles were included in the inter-observer test. To measure illuminance and luminance at the workplace a Hagner (Sweden) universal type S1 light meter was used. Viewing angles at the computer display were measured with the help of a protractor.

An inter-observer test included an interview part with the subject, who was asked about the operator's knowledge of how to adjust the chair height, how to adjust the armrest, how to adjust the backrest, etc.

3.5. Procedure

3.5.1. Questionnaire

The original questionnaire was filled in at each company during working time and took about 35 min to answer. Each questionnaire was put in an envelope that was either sent back to the project group, or was collected immediately after it was answered.

At retest—2 to 4 weeks later—a copy of the original questionnaire was filled in and mailed to the project group. This time period was considered to be long enough for the responders to forget their answers to the original questionnaire.

3.5.2. Ergonomic checklist

The inter-observer reliability of 44 selected variables in the ergonomic checklist was tested by two experienced and trained ergonomists, who made the observations, measurements, and interview coding independently of each other. The observations of the working postures were carried out simultaneously by the ergonomists, but the different measurements were not made at exactly the same time, because there was only one measuring instrument of each kind. The time between the two measurements ranged from a few minutes to a maximum of 30 min. The interviews were made by one ergonomist while the other one was standing beside and listening to the operator's answers.

Evaluation of the working postures was carried out on the most common posture during the observation period (assessed through observation), representing a typical working situation. The working posture in the neck was evaluated when the operator was looking at the screen or the keyboard. The working posture in the shoulder, wrist, and lower back was evaluated when the operator was using the input device, and if this was not possible the working posture was evaluated when the operator was using the keyboard.

[3] Retrieved January 19, 2006.

The items that were investigated, in the interview part, were how the chair and desk could be adjusted. The operator was asked what different adjustments could be made to the chair and the table. After this the operator was given the task of adjusting the height to what he/she believed was the optimal or best height. That height was measured and compared with the height that the ergonomist estimated as optimal or best.

4. STATISTICAL ANALYSIS

The distribution of responses in the original and retest questionnaires was described by the minimum, the 10th, 50th and 90th percentiles, and the maximum. The reliability of the questions, and the measurements were analysed by calculating Pearson's correlation coefficient for variables on ratio and interval level. Spearman's correlation coefficient was calculated for variables on ordinal level and Cohen's Kappa coefficient was calculated for variables on nominal level. The percentage agreement was calculated as a complement to Cohen's Kappa. The following categorization has been suggested for Pearson (*P*) and Spearman (*S*) correlation: *high* reliability ≥.90, *good* reliability .80–.89, *fair* reliability .70–.79, and *poor* reliability <.70 [7].

To simplify the presentation of data and to accommodate to the classification of Kappa values, we used three intervals for reliability: *high* reliability ≥.90, *fair* to *good* reliability .70–.89, and *poor* reliability <.70. When Kappa was used for evaluation of the reliability, the following rules of thumb were used: *high* reliability >.75, *fair* to *good* reliability .40–.75, and *poor* reliability <.4 [8].

Calculations of reliability were considered not meaningful when there were fewer than 10 pair-wise comparisons. For several of the follow-up questions we observed that they had fewer than 10 pair-wise comparisons. Kappa statistics could not be calculated when the distribution of answers was too uneven [9].

Cronbach's α was used to analyse the internal consistency of the constructed indices. Cronbach's α is based on the average correlation of items within a test, if the items are standardized to a standard deviation of 1; or the average covariance among items on a scale, if the items are not standardized. Cronbach's α is considered to be satisfactory when alpha is ≥.7 [2, 4]. If Cronbach's α is over 0.9 this could be a sign that the index includes several almost identical questions and some of them could be unnecessary. We assumed that the items in the indices used were positively correlated with each other because to a certain extent they were measuring a common entity. If the internal consistency is low (α ≤ .6) this can be interpreted as the items having a limited connection with each other.

The internal consistency of indices was calculated from the original and the retest questionnaires. The test-retest reliability of indices between the two occasions was also calculated.

All statistical analyses were performed with SPSS version 11.5 [10].

5. RESULTS

In the questionnaire there were 31 questions with *high* reliability, 111 questions with *fair* to *good* reliability, and 144 questions with *poor* reliability (Table 2). Fifteen questions were disregarded because of fewer than 10 pair-wise comparisons. For 11 questions Cohen's Kappa could not be calculated because 98–100% of the observations were in one cell. A complete list of each question and its reliability is shown at www.arbetslivsinstitutet.se/pdf/CCBaselineQuestReliab.pdf[4]. The average value (and range) of Pearson's correlation coefficient was .61 (–.031–1.00), Spearman correlation coefficient .59 (.33–.77), and Cohen's Kappa .59 (–.028–1.00).

Questions about background variables (at the ratio and nominal levels), employment (at the ratio and nominal levels), working hours and

[4] Retrieved January 19, 2006.

TABLE 2. The Main Groups of Questions, Their Average Test-Retest Reliability Coefficients (and Range), Classification of Reliability, and Average Percentage Agreement for Variables at the Nominal Level (*n* = 57).

Variable (No. in the Questionnaire)	Reliability Average (Min–Max)	Classification of Reliability	Average Agreement (%)
Background questions			
ratio level (1, 3, 4, 9 b)	*P* = .92 (.74–1.00)	*High*	
nominal level (2, 5, 6, 7, 8, 9 a, 10)	*K* = .86 (.74–1.00)	*High*	93
Employment			
ratio level (13, 14, 15)	*P* = .87 (.74–.94)	*Fair* to *good*	
nominal level (11, 16 a, b)	*K* = .74 (.64–.85)	*Fair* to *good*	90
Working hours and remuneration			
ratio level (17, 18, 21, 22)	*P* = .85 (.69–.96)	*Fair* to *good*	
ordinal level (20)	S = .62	*Poor*	
nominal level (19 a–c, 23 a–n, 24 a–i)	*K* = .58 (–.028–1.00)	*Fair* to *good*	92
Duties			
ratio level (25 a–d2, 25 e, 26, 27 b–c, 28, 29 b, 29 c, 31 a–d, 32 a–c, 33)	*P* = .52 (.005–.92)	*Poor*	
nominal level (25 a–d, 27 a, 29 a, 30)	*K* = .54 (.32–.68)	*Fair* to *good*	83
Computer work, workplace design, disruption and technical support			
ratio level (34, 35, 36)	*P* = .55 (.18–.87)	*Poor*	
ordinal level (37, 38 a–n, 39, 40)	*S* = .57 (.18–.74)	*Poor*	
Social support, management			
ordinal level (41 a–o)	*S* = .59 (.39–.76)	*Poor*	
Development			
ratio level (43 b)	*P* = .81	*Fair* to *good*	
ordinal level (42)	*S* = .45	*Poor*	
nominal level (43 a)	*K* = .76	*High*	89
Psychological demands, lack of control, limited decision latitude			
interval level (44 a–m, 45 a–q)	*P* = .59 (.20–.77)	*Poor*	
Call logging, monitoring			
nominal level (46 a–b1–8, 47 a–b1–8)	*K* = .62 (.46–.82)	*Fair* to *good*	89
Stress, energy and tiredness			
ordinal level (48 a–t)	*S* = .67 (.43–.77)	*Poor*	
Sleep, winding down and recovery			
ratio level (49, 50 a–g, 51 a–c)	*P* = .64 (.51–.74)	*Poor*	
ordinal level (52 a–d)	*S* = .68 (.59–.75)	*Poor*	
Work/life balance			
ordinal level (53 a–j, 54, 57)	*S* = .55 (.33–.71)	*Poor*	
Health, problems (No. of days), work-related problems			
ratio level (58 a2–f2, 59 a2–m2)	*P* = .58 (–.031–.87)	*Poor*	
nominal level (58 a–f, 58 a3–f3, 59 a–s, 59 a3–m3)	*K* = .56 (.18–1.0)	*Fair* to *good*	86
Measures taken to reduce problems, problems affecting working capacity			
ratio level (60 d2, 62 a–f)	*P* = .59 (–.50–.85)	*Poor*	
ordinal level (61 a–l)	*S* = .50 (.37–.67)	*Poor*	
nominal level (60 a–m)	*K* = .33 (–.067–.72)	*Poor*	74
Reported work-related injuries, sick-leave, eye test, glasses, summary of current health			
ratio level (64 b, 65 b)	*P* = .52 (.28–.75)	*Poor*	
ordinal level (67)	*S* = .66	*Poor*	
nominal level (63, 64 a, 64 c, 65 a, 66 a–c)	*K* = .66 (.36–.78)	*Fair* to *good*	88

Notes. *P*—Pearson's correlation coefficient, *S*—Spearman's correlation coefficient, *K*—Kappa coefficient.

remuneration (at the ratio and nominal levels), duties (at the nominal level), development (at the ratio and nominal levels), call logging and monitoring (at the nominal level), health (at the nominal level), reported and work-related injuries, etc., (at the nominal level) had *high* or *fair* to *good* reliability. Questions from the area of working hours and remuneration (at the ordinal level), duties (at the ratio level), computer work, workplace design, disruption in the computer system and technical support (at the ratio and ordinal levels), social support and support from the manager (at the ordinal level), development (at the ordinal level), psychosocial demands, lack of control (at the interval level), stress, energy, tiredness (at the ordinal level), sleep, winding down and recovery (at the ratio and ordinal levels), work/life balance (at the ordinal level), health problems (at the ratio level), measures taken to reduce the problems (at the ratio, ordinal and nominal levels) and reported and work-related injuries (at the ratio and ordinal levels) had *poor* reliability.

There was a lower proportion of variables on nominal level classified as having poor reliability compared with variables on ratio, interval, and ordinal levels (Table 3).

Twelve indices in the original questionnaire were classified as having satisfactory internal consistency ($\alpha \geq .7$) and four indices were classified as having low internal consistency ($\alpha \leq .6$) (Table 4). Indices with *satisfactory* internal consistency were comfort of sound, lighting and air quality, comfort of furniture and equipment, social support, support from supervisor, psychological demands, lack of control, cognitive demands, stress, positive work, feeling worn out, anxiousness, and psychosomatic symptoms. Indices with *low* internal consistency were limited decision latitude, time pressure, emotional demands, and energy.

Indices with *fair* to *good* test-retest reliability were comfort of furniture and equipment, psychological demands, limited decision latitude, lack of control, cognitive demands, positive work, feeling worn out, anxiousness, and psychosomatic symptoms. Indices with *poor* test-retest reliability were comfort of sound, lighting and air quality, social support, support from supervisor, time pressure, emotional demands, stress, and energy.

In the checklist there were 11 variables—out of 44—with *high* reliability, 16 variables with *fair* to *good* reliability, and 11 variables with *poor* reliability (Tables 5 and 6). Among the variables

TABLE 3. Variables in the Questionnaire at the Ratio, Interval, Ordinal, and Nominal Levels Classified into *High*, *Fair* to *Good* or *Poor* Test-Retest Reliability, or Not Classified Due to Fewer Than 10 Pair-Wise Observations or 98–100% of the Variables in One Cell

	Reliability											
	High		*Fair* to *good*		Poor		98–100% in One Cell		<10 Pair-Wise Observations		Total	
Scale of Measurement	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Ratio or interval level (Pearson's)	9	9	30	29	62	60	—	—	2	2	103	100
Ordinal level (Spearman's)	0	0	19	23	63	76	—	—	1	1	83	100
Nominal level (Kappa)	22	17	62	49	19	15	11	9	12	10	126	100
Total	31	10	111	36	144	46	11	4	15	5	312	100

TABLE 4. Internal Consistency (Cronbach's α) of Indices From the Questionnaire at Original and Retest Rounds and Test-Retest Reliability of Indices Between the Two Occasions (*n* = 57)

Index	No. of Items	Scale	Cronbach's α		Reliability	Classification of Reliability
			Original Questionnaire	Retest Questionnaire		
Comfort of sound, lighting, and air quality	5	1–5, *very dissatisfied–very satisfied*	.69	.68	*S* = .48	*Poor*
Comfort of furniture and equipment	9	1–5, *very dissatisfied–very satisfied*	.88	.89	*S* = .84	*Fair* to *good*
Social support	7	1–6, *never–always*	.80	.75	*S* = .63	*Poor*
Support from supervisor	8	1–6, *never–always*	.94	.94	*S* = .66	*Poor*
Psychological demands	14	1–11, 0–100% of the working time	.84	.86	*S* = .71	*Fair* to *good*
Limited decision latitude	4	1–11, 0–100% of the working time	.64	.62	*S* = .70	*Fair* to *good*
Lack of control	7	1–11, 0–100% of the working time	.72	.70	*S* = .73	*Fair* to *good*
Cognitive demands	7	1–11, 0–100% of the working time	.81	.81	*S* = .70	*Fair* to *good*
Time pressure	3	1–11, 0–100% of the working time	.54	.38	*S* = .50	*Poor*
Emotional demands	3	1–11, 0–100% of the working time	.57	.65	*S* = .52	*Poor*
Stress	4	0–5, *not at all–very, very*	.90	.88	*S* = .68	*Poor*
Energy	4	0–5, *not at all–very, very*	.62	.71	*S* = .65	*Poor*
Positive work	16	1–11, 0–100% of the working time	.68	.62	*S* = .76	*Fair* to *good*
Feeling worn out	3	0–5, *not at all–very, very*	.95	.95	*S* = .79	*Fair* to *good*
Anxiousness	4	0–5, *not at all–very, very*	.90	.88	*S* = .80	*Fair* to *good*
Psychosomatic symptoms	5	1–11, 0–100% of days	.80	.79	*S* = .86	*Fair* to *good*

Notes. S—Spearman's correlation coefficient.

on ratio and ordinal level in the checklist, measurements of actual desk height and seat height adjusted by the operator had *high* reliability (Table 5). Viewing angle, some of the luminance variables, e.g., risk for glare in the field of vision and luminance in the peripheral working area had *poor* reliability. Three variables were disregarded due to few pair-wise observations ($n < 10$). For three variables Cohen's Kappa could not be calculated when 98–100% of the observations were in one cell.

Among the variables on the nominal level, function of curtains, awnings, etc., and neck posture when looking at the screen and the operator's knowledge concerning how to adjust the armrest had *high* reliability (Table 6). Variables such as visible reflections on the desk, risk of reflections on the desk, reflections on the screen, the keyboard positioned within forearm's length, and shoulder width were classified as having *poor* reliability.

TABLE 5. Inter-Rater Reliability of Variables at the Ratio and Ordinal Levels in the Ergonomic Checklist for Measurement 1 and 2 (*n* = 58)

Variable (No. in the Ergonomic Checklist)	Scale	Measurement 1 Min/P10/**P50**/P90/Max	Measurement 2 Min/P10/**P50**/P90/Max	Reliability	Classification of Reliability
Actual desk height (66)	cm	66/71/**77**/84/115	67/69/**79**/104/114	*P* = .98	*High*
Seat height adjusted by the operator (124)	cm	38/46/**51**/55/67	41/47/**51**/56/51	*P* = .96	*High*
Luminance: inner working area (VDU screen) (36 a)	cd/m^2	5/35/**81**/150/210	29/49/**108**/175/220	*P* = .89	*Fair* to *good*
Luminance: outer working area (36 b)	cd/m^2	3/11/**44**/85/190	4/15/**50**/91/120	*P* = .71	*Fair* to *good*
Luminance: peripheral working area, lightest surface (36 c)	cd/m^2	10/35/**412**/900/4500	35/45/**355**/720/4000	*P* = .83	*Fair* to *good*
Actual seat height (52)	cm	42/47/**52**/57/61	46/48/**52**/56/59	*P* = .87	*Fair* to *good*
Shoulder extension (115 a)	4 categories	1/2/**2**/4/4	1/2/**2**/4/4	*S* = .89	*Fair* to *good*
Shoulder abduction (115 b)	3 categories	1/1/**2**/3/3	1/2/**2**/3/3	*S* = .71	*Fair* to *good*
Shoulder inward rotation (115 c)	5 categories	1/2/**3**/4/5	1/2/**3**/4/5	*S* = .89	*Fair* to *good*
Optimal seat height (125)	cm	38/45/**49**/53/69	40/45/**48**/51/56	*P* = .81	*Fair* to *good*
Luminance, risk of glare in the field of vision, ceiling (37 b)	cd/m^2	5/30/**602**/1200/5600	47/150/**454**/830/1200	*P* = .59	*Poor*
Luminance: peripheral working area, darkest surface (36 d)	cd/m^2	2/2/**31**/51/300	0/15/**24**/32/100	*P* = .57	*Poor*
Luminance, risk of glare in the field of vision, ceiling luminaries (37 a)	cd/m^2	5/30/**602**/1200/5600	47/150/**454**/830/1200	*P* = .30	*Poor*
Viewing angle to the top edge of the screen (88 a, 88 a1)	°	0/1/**5**/10/17	1/1/**5**/10/15	*P* = .61	*Poor*
Viewing angle to the bottom edge of the screen (88 b, 88 b1)	°	6/15/**22**/32/40	11/15/**21**/27/34	*P* = .54	*Poor*
Neck extension (114 a)	4 categories	1/2/**2**/2/3	1/2/**2**/2/3	*S* = .50	*Poor*
Neck rotation (114 c)	5 categories	1/1/**1**/1/3	1/1/**1**/1/3	*S* [1]	*Poor*

Notes. Min—minimum, P10—10th percentile, P50—50th percentile, P90—90th percentile, Max—maximum; *P*—Pearson's correlation coefficient, *S*—Spearman's correlation coefficient; 1—variables were disregarded due to few observations ($n < 10$).

TABLE 6. Inter-Rater Reliability of Variables at the Nominal Level in the Ergonomic Checklist; Median, Frequencies, Range and the Reliability of Classifications and Percentage Agreement Between Measurements 1 and 2 (*n* = 58)

Variable (No. in the Ergonomic Checklist)	Scale	Measurement 1 Median (Frequencies %) (Range)	Measurement 2 Median (Frequencies %) (Range)	Reliability	Classification of Reliability	Agreement (%)
Curtains, blinds, awnings, and/or solar film function (25 b)	*yes/no*	*yes* (*f* = 99) (*yes/no*)	*yes* (*f* = 97) (*yes/no*)	*K* = 1.0	*High*	100
Backrest height (57)	3 categories	2 (*f* = 63) (1–3)	2 (*f* = 71) (1–3)	*K* = .77	*High*	89
Is the control device positioned within forearm's length and shoulder width (106)	*yes/no*	*no* (*f* = 83) (*yes/no*)	*no* (*f* = 88) (*yes/no*)	*K* = .84	*High*	97
Neck posture (looking at the screen or keyboard) (114)	2 categories	1 (*f* = 97) (1–2)	1 (*f* = 93) (1–2)	*K* = 1.0	*High*	100

Table 6. (*continued*)

Variable (No. in the Ergonomic Checklist)	Scale	Measurement 1 Median (Frequencies %) (Range)	Measurement 2 Median (Frequencies %) (Range)	Reliability	Classification of Reliability	Agreement (%)
Shoulder joint posture (using control device or keyboard) (115)	2 categories	1 (*f* = 72) (1–2)	1 (*f* = 59) (1–2)	*K* = .96	*High*	98
Operator's knowledge of how to adjust chair height (123 a)	3 categories	2 (*f* = 97) (1–3)	2 (*f* = 95) (1–3)	*K* = .79	*High*	98
Operator's knowledge of how to adjust armrest (123 b)	3 categories	2 (*f* = 57) (1–3)	2 (*f* = 59) (1–3)	*K* = 1.0	*High*	100
Operator's knowledge of how to adjust seat depth (123 d)	3 categories	2 (*f* = 43) (1–3)	2 (*f* = 39) (1–3)	*K* = .89	*High*	93
Operator's knowledge of how to adjust tilt function (123 e)	3 categories	2 (*f* = 59) (1–3)	2 (*f* = 71) (1–3)	*K* = .84	*High*	93
Curtains, blinds, awnings and/or solar film are used (25 a)	*yes/no*	*yes* (*f* = 99) (*yes/no*)	*yes* (*f* = 97) (*yes/no*)	*K* = .66	*Fair* to *good*	98
Main source of disturbing noise (43)	4 categories	3 (*f* = 88) (1–4)	3 (*f* = 93) (1–4)	*K* = .48	*Fair* to *good*	93
Backrest is narrow/wide (58)	3 categories	1 (*f* = 77) (1–3)	1 (*f* = 88) (1–3)	*K* = .62	*Fair* to *good*	91
Display screen is positioned so that operator is exposed to glare from daylight (90)	*yes/no*	*no* (*f* = 75) (*yes/no*)	*no* (*f* = 76) (*yes/no*)	*K* = .49	*Fair* to *good*	81
Reflections from daylight in VDU screen (91)	*yes/no*	*no* (*f* = 78) (*yes/no*)	*no* (*f* = 79) (*yes/no*)	*K* = .58	*Fair* to *good*	86
Craned neck (114 b)	*yes/no*	*no* (*f* = 90) (*yes/no*)	*no* (*f* = 93) (*yes/no*)	*K* = .73	*Fair* to *good*	96
At least half of the forearm and/or elbow are supported (117)	*yes/no*	*yes* (*f* = 76) (*yes/no*)	*yes* (*f* = 78) (*yes/no*)	*K* = .64	*Fair* to *good*	88
Operator's knowledge of how to adjust backrest (123 c)	3 categories	2 (*f* = 65) (1–3)	2 (*f* = 64) (1–3)	*K* = .70	*Fair* to *good*	86
There are visible reflections on the desk (38)	*yes/no*	*no* (*f* = 82) (*yes/no*)	*no* (*f* = 84) (*yes/no*)	*K* = .34	*Poor*	83
There is risk of reflections on the desk (39)	*yes/no*	*no* (*f* = 67) (*yes/no*)	*no* (*f* = 71) (*yes/no*)	*K* = .25	*Poor*	67
Are there reflections on the screen from luminaries, shiny surface, etc. (86)	*yes/no*	*no* (*f* = 58) (*yes/no*)	*no* (*f* = 81) (*yes/no*)	*K* = .21	*Poor*	72
Viewing angle, below/above the horizontal plane (top of the edge of the screen) (88 a1)	2 categories	1 (*f* = 79) (1–2)	1 (*f* = 79) (1–2)	*K* = .29	*Poor*	55
The keyboard positioned within forearm's length and shoulder width (100)	*yes/no*	*no* (*f* = 51) (*yes/no*)	*yes* (*f* = 53) (*yes/no*)	*K* = .071	*Poor*	53
Task lighting adjustable direction (33 a)	*yes/no*	*yes* (*f* = 98) (*yes/no*)	*yes* (*f* = 100)	*K* [1]		
Task lighting adjustable lighting level (33 b)	*yes/no*	*no* (*f* = 100) (*yes/no*)	(*yes/no*)	*K* [1]		
Sufficient room for working material (72)	*yes/no*	*yes* (*f* = 89) (*yes/no*)	*no* (*f* = 100) (*yes/no*)	*K* [2]		100
Possible to support forearms, using keyboard on the desktop (73)	*yes/no*	*yes* (*f* = 99) (*yes/no*)	*yes* (*f* = 98) (*yes/no*)	*K* [3]		98
Possible to support forearms, using control device on the desktop (74)	*yes/no*	*yes* (*f* = 98) (*yes/no*)	*yes* (*f* = 100) (*yes/no*)	*K* [2]		100

Notes. *K*—Kappa coefficient; 1—variables were disregarded due to few observations ($n < 10$), 2—*K* could not be calculated when 100% of the observations were in one cell, 3—*K* could not be calculated when 98% of the observations were in one cell.

6. DISCUSSION

6.1. Reliability of the Questions in the Questionnaire

The test-retest reliability of the questions seems to be *fair* to *good* or better in 142 (46%) of the total of 312 questions. Variables on ratio, interval, and nominal levels had variables represented in all three categories of reliability, whereas no variables on the ordinal level were classified as having *high* reliability. Additionally, variables on the ordinal level had the smallest proportion of variables classified as having *fair* to *good* reliability. Only 15% of the variables on the nominal level were classified as having *poor* reliability compared with 60% of the variables on ratio or interval levels and 76% of the variables on the ordinal level.

Wikman [11] has shown that the reliability in survey questions concerning working environment varies considerably between different questions and that for many questions reliability is bad over time. Further, he has shown that the reliability of questions about facts is better than of questions involving evaluation and judgment. This has also been confirmed in this study; for example, background questions (average $P = .92$) showed better reliability compared with questions that were based on evaluation and feelings about conditions, e.g., social support and psychosocial questions (average $P = .59$).

Wictorin et al. [12] found that the reliability of questions regarding physical exposure was low when the answers to questions were accumulated in one part of the scale. Examples of questions in this study where the answers were accumulated in one part of the scale were those about psychological demands, lack of control, and limited decision latitude (No. 44 c–f, h–i, k, m; and No. 45 a, c–e, g–h, k–m, q). These questions had *poor* reliability. In the Wictorin et al. study the Kappa coefficient reached values exceeding .40, only if the lower parts of the scale were dichotomised, e.g., differentiating between *not exposed* and *exposed*. In our study we found an average value of Cohen's Kappa of .59.

Franzblau et al. [13] found that test-retest reliability of the questionnaire used to elicit demographic information, medical history, participation in exercise, and information on musculoskeletal symptoms among industrial workers appeared to be *good* to *excellent* in most instances (average Kappa = .76). This suggests that most variables of the questionnaire were reliable and suitable for use in epidemiological studies. In our study we found that symptoms involving the neck, shoulders, upper arms, hand, and fingers appeared to have *fair* to *good* reliability (average Kappa = .62). Several other studies have used the test-retest method as a method to evaluate the reliability of questions. Leijon et al. [14] found a test-retest agreement (weighted Kappa) for questions about physical workload varying from .74 to .92. Salerno et al. [15] found that Kappa values for questions about symptoms varied between .60 and .89. Booth-Jones et al. [16] found that Kappa values for questions about musculoskeletal symptoms and work history ranged between .46 and .77.

6.2. Internal Consistency and Reliability of the Indices

Twelve out of the total of 16 indices were classified as having *satisfactory* internal consistency and four indices were classified as having *low* internal consistency. Examples of indices with *satisfactory* internal consistency were comfort of furniture and equipment, social support, lack of control, anxiousness, feeling worn out, and psychosomatic symptoms. Indices with *low* internal consistency were limited decision latitude, time pressure, emotional demands, and positive work. The internal consistency for time pressure was the index that had changed most, compared with the original questionnaire. Furthermore, indices for energy and emotional demands changed substantially compared to the original questionnaire. Indices with *low* internal consistency may include variables that may change independently of each other.

Nine indices were classified as having *fair* to *good* test-retest reliability. Examples of indices with *fair* to *good* reliability were psychosomatic symptoms, comfort of furniture and equipment, anxiousness, and feeling worn out. Seven indices were classified as having *poor* test-retest reliability.

Examples of indices with *poor* reliability were comfort of noise, lighting, time pressure, emotional demands, and energy. Seven of 11 indices with *satisfactory* internal consistency had *fair* to *good* test-retest reliability. It is reasonable to assume that indices with *satisfactory* internal consistency are more likely to get good test-retest reliability. Balogh et al. [17] studied the reliability of indices regarding different kinds of self-rated exposure, like work postures and biomechanical load, and found that test-retest reliability was *good* or better (weighted Kappa >.6) for all 10 included indices. For 5 of the indices Cronbach's alpha was >.8.

6.3. Reliability of Assessments in the Ergonomic Checklist

A majority of the assessments, 27 variables out of 44, in the ergonomic checklist were classified as having *fair* to *good* or higher reliability. Examples of variables that had *high* reliability were actual desk height and seat height adjusted by the operator, the function of curtains, etc., neck position, and the operator's knowledge concerning how to adjust the armrest. There were 10 variables that were classified as having *poor* reliability. Examples of variables that had *poor* reliability were viewing angle, glare and luminance, visible reflections on the desk, risk of reflections on the desk, reflections on the screen, the keyboard positioned within forearm's length, and shoulder width. Some of these variables could be affected by instability and the fact that the ergonomists observed the operators from different angles**.** Stavem et al. [18] found for pairs of observers, the inter-observer agreement of audit of quality of radiology requests and reports was generally high; however, the corresponding Kappa values were low with only 14 of 90 ratings >.6 and 6 > .8. Sagaram et al. [19] found a poor inter-rater agreement for 10 out of the 22 quality criteria applied to online health information, and 15 out of the 22 had a Kappa value >.6.

6.4. Factors Affecting the Test-Retest Reliability

The correlation between the two measurements could be affected by lack of stability of the measured variables, type of questions, type of scale categories and differences in the distribution of answers, reactivity and memory effects, differences in response rate to the questions, and by random factors. Our intention was only to measure the random factor in this study. A major problem is to make a distinction between these sources of error.

The results from this study show the difficulty to reach acceptable reliability in survey studies. The questions in the questionnaire reflect several conditions that are variable, leading to notions and evaluations that could be under gradual development or change. Working conditions, health, and wellbeing could therefore have changed between the test and retest measurements. A *low* test-retest correlation may not indicate that the reliability, per se, of the test is *poor* but may, instead, signify that the underlying physical reality itself has changed. Test-retest correlations can underestimate the degree of reliability in measurements over time by interpreting true changes as measurement instability.

There are a number of question areas in this study that involve evaluation and judgement that concern the individual's perceptions. We used a test-retest method, and individual answers to questions may have been related to feelings and perceptions, which may be affected by real changes in working life as well as by mood [20]. There could have been questions that were difficult or even impossible to answer, e.g., when a subject is asked if a present health problem is work-related. It can be questioned whether a subject is really able to answer this question correctly. He or she might not have the correct knowledge needed to decide whether a problem is work-related or not. Questions that concern concrete conditions in the operator's life might be easier to formulate and might be understood as neutral and easier to respond to. This could lead to good and unambiguous answers in the survey. Other questions that concern subject estimations or attitudes are more difficult to formulate in an unambiguous and neutral way. They could also be perceived as difficult to answer or too personal, which might lead to more uncertain answers. Questions may also be interpreted in ways that were not intended.

A possible weakness of this survey is that the response scales may have been too detailed in relation to the subject's rating ability. It may be too difficult to estimate time consistently in proportions of a typical working day and frequencies in times per hour. It could also be a problem if the response scale covers the wrong area leading to a concentration of the answers in one end of the scale. For example, in this study, answers to questions concerning job demands were concentrated in the upper part of the scale. How detailed the scale should be depends on the purpose for which a specific question is intended. In epidemiological analyses, where relative risks are studied, a rough quantification of exposure may be enough, compared with a situation when the purpose is to describe the working conditions or to evaluate measures in the working environment. A detailed scale also has advantages. If a greater initial number of categories is later combined to fewer categories, provided the response frequencies become more evenly distributed, the reliability may be better than if the number of categories is low from the beginning. We also found that reliability increased with the number of steps in the scale, e.g., when we used an 11-step response scale for questions about job demands (No. 44 a–m), the reliability coefficient was between .41 and .72. We then reduced the response scale to five and three steps, and the reliability coefficient decreased to between .38 and .64, and further to between .12 and .58, respectively. A more detailed scale favors high reliability, if the answers are distributed over all categories in the scale.

Reactivity is a problem that affects test-retest correlations. It refers to the fact that sometimes the process of measuring a phenomenon can induce a change in the experience of the phenomenon itself and will thus deflate the reliability estimate. The first measurement could have affected the subjects by making it easier to interpret the questions the second time they received the questionnaire or by increasing awareness of the phenomenon.

Another methodological problem is that if the time interval between measurements is relatively short, the subjects might remember their earlier responses and the questions will appear more reliable than they actually are. Memory effects lead to inflated reliability estimates [21]. It is recommended that the two tests be administered about 2 weeks apart, thus allowing for day-to-day fluctuations in a person [21]. Thus, the test-retest method will often provide a substantial overestimate of what would be obtained from the split-half method [22].

Differences in response rates to questions may also affect the reliability coefficient. If there are fewer than 10 pair-wise observations, a reliability coefficient will not be meaningful.

The measurement error of the different instruments used is another possible source of error. Are they stable or are they sensitive to external influences? The Hagner (Sweden) universal light meter was calibrated before each measurement was started, which probably minimized the measuring error.

6.5. Factors Affecting the Inter-Observer Reliability

The *poor* reliability for some items in the ergonomic checklist may be due to rapid changes of some variables, e.g., viewing angle and work postures, and the two observers might not have observed these variables at exactly the same time. Poor reliability was also found for measurements concerning existing visual reflections on the desk and reflections on the screen from different light sources. One explanation for the poor reliability of these variables could be that the ergonomists did not observe these variables from the same angle.

6.6. Problems With Reliability Coefficients

A problem with the reliability coefficients is that its value, as well as the level of agreement between the observations, depends on the frequency distribution of the variable in the sample. The reliability coefficients increase the more evenly distributed and the greater the range of responses. If the answers are aggregated in one part of the scale the correlation becomes low. A variable with a low Kappa value could have a high percentage agreement. Thus, it may be misleading to compare the reliability between variables with quite different distributions. Therefore it is necessary to look at other ways to describe reliability.

One weakness of Cohen's Kappa is that the coefficient could not be calculated when 98–100% of the observations were in one cell or when there were too few observations ($n < 10$). In some cases the number of observations was low and Kappa could not be calculated.

A possible bias from reactivity is if there is a systematic change in response at retest. This does not influence the reliability coefficient but the intercept, which was not studied in the present investigation. However, there is no reason to believe that there would be systematic differences between the test and the retest.

7. CONCLUSIONS

About half of the questions in the self-administrated questionnaire in this study could be classified as having *fair* to *good* or higher test-retest reliability. These questions can be recommended in further analyses. Other questions should be used with care.

The ergonomic checklist used in this study appears to have a majority of variables that could be classified as having *fair* to *good* or higher inter-rater reliability.

Low reliability does not necessarily indicate that reliability of the test, per se, is low but may signify that the conditions measured vary over time or that answers are aggregated in one part of the scale.

REFERENCES

1. Fleiss JL. The design and analysis of clinical experiments. New York, NY, USA: Wiley; 1986.
2. Dzissah JS, Karwowski W, Rieger J, Stewart D. Measurement of management efforts with respect to integration of quality, safety, and ergonomics issues in manufacturing industry. Human Factors and Ergonomics in Manufacturing. 2005;15(2): 213–32.
3. Kerlinger FN, Lee HW. (2000). Foundations of behavioural research. 4th ed. Fort Worth, TX, USA: Harcourt College; 2000.
4. Bland JM, Altman DG. Statistic notes: Cronbach's alpha. BMJ. 1997;314:572.
5. Norman K, Nilsson T, Hagberg M, Wigaeus Tornqvist E, Hagman M, et al. Working conditions and health among female and male employees at one call centre in Sweden. Am Jour Ind Med. 2004; 46(1):55–62.
6. Tengblad P, Wiberg A, Herrman L, Backström M. Hållbart arbete i informationssamhället; Call Centre i utveckling (Report No. VR 2002:7). Stockholm, Sweden: Vinnova; 2002.
7. Currier DP. Elements of research in physical therapy. 2nd ed. London, UK: Williams & Wilkins; 1984.
8. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY, USA: Wiley; 1981.
9. Maclure M, Willett W. Misinterpretation and misuse of the Kappa statistic. Am J Epidemiol. 1987;126(2):161–9.
10. SPSS base system user's guide. Chicago, IL, USA: SPSS; 1990.
11. Wikman A. Developing social indicators—a survey method illustrated with examples of the working environment [doctoral dissertation]. Stockholm University: Stockholm Statistics Urval; 1991. In Swedish.
12. Wictorin C, Karlqvist L, Winkel J. Validity of self-reported exposures to work postures and manual handling. Scand J Work Environ Health. 1993;19:208–14.
13. Franzblau A, Salerno DF, Armstrong TJ, Werner RA. Test-retest reliability of an upper-extremity discomfort questionnaire in an industrial population. Scand J Work Environ Health. 1997;23(4):299–307.
14. Leijon O, Wiktorin C, Härenstam A, Karlqvist L, MOA Research Group. Validity of a self-administrated questionnaire for assessing physical work loads in a general population. J Occup Environ Med. 2002; 44(8):724–35.
15. Salerno DF, Franzblau A, Armstrong TJ, Werner RA, Becker MP. Test-retest reliability of the upper extremity questionnaire among keyboard operators. Am J Ind Med. 2001;40(6):655–66.
16. Booth-Jones AD, Lemasters GK, Succop P, Atterbury MR, Bhattacharya A. Reliability of questionnaire information measuring musculoskeletal symptoms and work

histories. Am Ind Hyg Assoc J. 1998; 59(1):20–4.

17. Balogh I, Orbaek P, Winkel J, Nordander C, Ohlsson K, Ektor-Andersen J. Questionnaire-based mechanical exposure indices for large population studies—reliability, internal consistency and predictive validity. Scand J Work Environ Health. 2001;27(1):41–8.
18. Stavem K, Foss T, Botnmark O, Andersen OK, Erikssen J. Inter-observer agreement in audit of quality of radiology request and reports. Clin Radiol. 2004;59(11):1018–24.
19. Sagaram S, Walji M, Meric-Bernstam F, Johnson C, Bernstam E. Inter-observer agreement for quality measures applied to online health information. MedInfo. 2004;11(Pt. 2):1308–12.
20. Barroso M, Wilson JR. HEDOMS—Human error and disturbance occurrence in manufacturing systems: toward the development of an analytical framework. Human Factors and Ergonomics in Manufacturing. 1999;9(1):87–104.
21. Nunnally JC. Psychometric theory. New York, NY, USA: McGraw-Hill; 1978.
22. Nunnally JC. Educational measurement and evaluation. New York, NY, USA: McGraw-Hill; 1964.