INTERNATIONAL STANDARD

ISO 10075-3

First edition
2004-08-15

Ergonomic principles related to mental workload —

Part 3: Principles and requirements concerning methods for measuring and assessing mental workload

Principes ergonomiques relatifs à la charge de travail mental —

Partie 3: Principes et exigences concernant les méthodes de mesurage et d'évaluation de la charge de travail mental

Reference number
ISO 10075-3:2004(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO 10075-3 was prepared by Technical Committee ISO/TC 159, *Ergonomics*, Subcommittee SC 1, *Ergonomic guiding principles*.

ISO 10075 consists of the following parts, under the general title *Ergonomic principles related to mental workload*:

— *Part 1: General terms and definitions*

— *Part 2: Design principles*

— *Part 3: Principles and requirements concerning methods for measuring and assessing mental workload*

A Technical Report will accompany these parts to explain to non-experts the basic concepts and how to use these parts.

Introduction

This part of ISO 10075 specifies technical information relevant in the context of constructing, evaluating and choosing measurement instruments for assessing mental workload as defined and treated in ISO 10075 and ISO 10075-2. Familiarity with the concepts discussed in these two documents is required to understand the provisions of this part of ISO 10075.

Since mental workload is a part of the total workload, users of this part of ISO 10075 should also be familiar with the concepts and provisions presented in ISO 6385.

This part of ISO 10075 aims at providing information for the development of measurement instruments, about which specifications will be required to evaluate a given procedure with regard to its usability as a measuring instrument for assessing mental workload.

This part of ISO 10075 addresses requirements for instruments measuring different aspects of mental workload, but it does not specify which instruments should be used, e.g. psychological scaling or psychophysiological methods. The choice of which instruments to use can be facilitated by the provision of appropriate information.

Ergonomic principles related to mental workload —

Part 3: Principles and requirements concerning methods for measuring and assessing mental workload

1 Scope

This part of ISO 10075 establishes principles and requirements for the measurement and assessment of mental workload and specifies the requirements for measurement instruments. This part of ISO 10075 provides information for choosing appropriate methods and provides information on aspects of assessing and measuring mental workload to improve communication among the parties involved.

This part of ISO 10075 is intended for use mainly by ergonomic experts, for example, psychologists, occupational health specialists, and/or physiologists, with appropriate training in the theoretical background and usage of such methods, as well as in the interpretation of the results. They will find the information needed when developing or evaluating methods of mental-workload assessment.

Non-experts, e.g. employers, employees and their representatives, system managers and designers, and public authorities can find useful information for their orientation in the field of assessment and measurement of mental workload, e.g. what kinds of methods are available, which criteria are relevant in the evaluation of measurement instruments and what kind of information they should require and observe in deciding which instrument will be suitable for their purpose and which can be used.

NOTE A Technical Report on the terminology and use of this part of ISO 10075 will be available for further information for non-experts.

This part of ISO 10075 provides information on which to base a well-considered choice for an appropriate method in different situations. There are a large number of different methods available which are suitable for different purposes, situations and different levels of precision. There is a need for effective and efficient methods of measurement. The information provided in this part of ISO 10075 will allow users to evaluate the type of measurement approach most suitable for their specific purposes.

Conformance with the provisions of this part of ISO 10075 has to be provided by the documentation requirements.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 6385:2004, *Ergonomic principles in the design of work systems*

ISO 10075:1991[1)], *Ergonomic principles related to mental workload — General terms and definitions*

ISO 10075-2:1996, *Ergonomic principles related to mental workload — Part 2: Design principles*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 6385:2004, ISO 10075:1991 and ISO 10075-2:1996 and the following apply.

3.1
objectivity
degree to which results gained with an instrument are independent of the person administering the instrument, analysing and interpreting the data

3.2
reliability
degree of precision to which a method or instrument is able to measure what it measures

NOTE Reliability can be assessed as homogeneity, consistency or stability of measurement, or in the case of two or more raters, as inter-rater-reliability. Reliability is closely related to generalizability.

3.2.1
homogeneity
degree to which all parts or items of a measurement procedure measure the same characteristic

3.2.2
consistency
degree to which different parts or parallel forms of a measurement instrument lead to identical results, e.g. by dividing a scale into two or more parts or applying two or more parallel forms of an instrument

3.2.3
stability
degree to which a (usually time delayed) replication of a measurement procedure leads to identical results

3.2.4
inter-rater-reliability
degree to which two or more raters produce the same results in measuring the same characteristics

3.3
validity
degree to which a method or instrument is able to measure what it is intended to measure

NOTE Validity can be assessed via concurrent validation (e.g. by agreement of results with the results of a simultaneously applied procedure known to measure the intended aspect), criterion-related validation (e.g. by establishing a relation with a relevant criterion), or factorial validity (e.g. by demonstrating that a measurement procedure assesses specific facets of a construct).

3.4
sensitivity
degree to which a method or instrument is able to discriminate between different degrees of the object of measurement, e.g. different degrees of mental strain or fatigue

1) If revised, this International Standard will become ISO 10075-1.

3.5
diagnosticity
degree to which a method or instrument is able to discriminate between different kinds or sources of mental workload e.g. perceptual demands etc., or its effects, e.g. discriminating between fatigue, monotony, satiation or reduced vigilance

3.6
generalizability
degree to which an observed score (unit of measurement) can be generalized to a defined universe of situations (stress conditions) and/or population of workers

3.6.1
relative generalizability
degree to which rank ordering of workload conditions/people, e.g. from low to high, is replicable (as in decisions concerning relative positions)

3.6.2
absolute generalizability
degree to which an absolute level of stress/strain associated with specific work conditions/people without regard to the stress/strain associated with other conditions/people can be replicated (as in decisions concerning absolute values)

NOTE For more explanations concerning generalizability, see Annex A.

3.7
usability
extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use

[ISO 9241-11:1998, Definition 3.1]

NOTE In the case of this part of ISO 10075, a product is an instrument.

3.8
critical values
reference standard for the evaluation of measured scores and/or pass-fail decisions

NOTE The establishment of critical values presupposes that the scores of the measurement instrument in question allow absolute decisions with high precision. Furthermore, it has to be stated under which conditions the critical values are valid (for example, post-test measurements etc.)

4 Measurement and assessment of mental workload

4.1 General principles

Mental workload is not a unitary and not a unidimensional concept, so the assessment and measurement of mental workload cannot be a uniform procedure. There is no single best way to assess mental workload, since the most appropriate way to assess or measure mental workload will depend on the purpose of the assessment, which may require the assessment of different aspects of mental workload, the use of different techniques of measurement, and different degrees of precision.

Thus, the model of workload assessment used in this part of ISO 10075 has a three-dimensional structure. It takes into account

— *different aspects of mental workload*, e.g. mental stress, mental strain, mental fatigue, etc.,

— *different techniques of measurement*, e.g. task analysis, performance assessment, subjective ratings or psychophysiological measurement, and

— *different degrees of precision*, e.g. measurement at an orienting, screening or accurate level of measurement.

According to ISO 10075:1991, different aspects have to be differentiated: mental stress, mental strain, and the effects of mental strain within the individual. Assessment and measurement thus have to refer to these different steps in the stress-strain-effects process, i.e:

— assessment of work conditions producing mental stress, as in the design and evaluation of work system design;

— assessment and measurement of mental strain, produced by mental stress, e.g. in order to evaluate the tolerability of the strain;

— measurement of the effects of strain in the employee, e.g. fatigue, monotony, satiation or reduced vigilance;

which may affect her/his health and safety, well-being, performance and productivity.

Different purposes may require different approaches and/or degrees of precision, e.g. a risk analysis for hazardous technologies as opposed to a survey where the intention is to give a general overview over different work systems. This part of ISO 10075 does not prescribe which level of precision has to be used, since this is a matter of the purpose and conditions of measurement, depending, for example, on legal requirements or agreed regulations, or cost-benefits considerations.

Methods for the measurement and assessment of mental workload, independent of the technique to be used, will be classified in this part of ISO 10075 according to three, i.e. low, intermediate and high levels of precision:

— level 1: for accurate measurement purposes. The purpose of the mental-workload assessment at this high precision level is to obtain reliable and valid information about the nature of the source of under-overload in order to optimise work conditions. The methods at this level will most probably be applicable only by specialists, e.g. psychologists, ergonomists, occupational health specialists, and/or physiologists, with appropriate training in the theoretical background and usage of these methods, as well as in the interpretation of the results.

— level 2: for screening purposes. Measurement at this level represents an intermediate level of precision, and is most often used for screening purposes, which require more precision than orienting measurements, e.g. where problems of mental workload can be anticipated or where causes of inappropriate workload have to be identified. Procedures to be used at this level have higher reliabilities, and demonstrated validities, and should be able to indicate if corrective measures should be taken.

— level 3: for orienting purposes. Methods at this level allow the user to gather information about mental workload at a low level of precision. They provide general information about work conditions, subjective and psychophysiological states of the worker with respect to mental workload, without being excessively resource intensive. The information at this level should make it possible to prevent negative effects by making management decisions at an operational level such as to change work tasks and/or methods and conditions. Measurement at this level is usually restricted to orienting measurements, e.g. relating to rough measures (with moderate levels of reliability, validity, etc.) of task analysis, subjective ratings of acceptability of work conditions and subjective states with regard to all aspects of mental workload.

In specifying for which level of precision an instrument can be used, the object of measurement should be indicated. It should be stated whether the object of measurement is a person or a situational characteristic. If the intended object of measurement is the individual, psychometric properties have to be assessed and specified for the assessment of people. If the intended object of measurement is a situational characteristic, psychometric properties shall relate to assessing conditions. This will be demonstrated in more detail in the context of testing generalizability (see 4.2.7).

In specifying for which level of precision an instrument can be used, a differentiation is also required depending on whether the measurement is based on single scores/observations or mean scores/observations.

All assessment procedures will inevitably carry a measurement error which can be reduced by averaging over a number of scores/observations. An instrument which reaches the orienting level requirements for single scores/observations may attain screening level requirements for average scores/observations if a sufficient number of scores is used. The required number of scores/observations for attaining a higher level of precision shall thus be specified.

An instrument claiming to assess mental workload shall have demonstrated validity in the assessment of those aspects for which validity is claimed, and the domains of measurement, for which validity is claimed, shall be clearly stated (e.g. mental fatigue or monotony). If validity is claimed for more than one aspect, documentation of the validity shall include evidence for each field of measurement.

Different techniques may be used to assess mental workload, with some techniques being more suitable for some domains of measurement than others. In particular, the following techniques can be applied:

— physiological measurements: these methods provide information about physiological states of employees under given work conditions;

— subjective scaling: these methods provide information on how employees subjectively assess different aspects of mental workload at their work stations, e.g. using psychometric scales, and how they feel about their work conditions;

— performance assessment: these methods offer the possibility to evaluate human mental and psychomotor performance under given work conditions, e.g. in order to assess decrements or variations in performance due to the effects of mental workload;

— job and task analysis: these methods assess task elements, physical and psychosocial work conditions, environmental conditions and the organization of the work process as sources of mental workload.

Different methods will be required to achieve different levels of reliability, e.g. a short questionnaire may be sufficient for an orienting level measurement, whereas in order to verify that a system design does not lead to monotony, scales with sufficient reliability for that purpose will be required. In order to safeguard safety-critical or hazardous systems against any negative effects of mental workload, the highest reliabilities and validities will be required. Where such instruments are not yet available, they should be developed. Until then, those methods with the highest available psychometric criteria shall be used for such purposes. Precision is not determined by the measurement technique itself, but by the development, the psychometric properties, and the adequate application of the method or instrument.

If measurement methods have to be used which do not fulfil the requirements, special expertise in the field of mental workload and its measurement will be required

— to assess the risks associated with using a less than perfect instrument, and

— to achieve a sound evaluation of the results.

However, the selection of an appropriate instrument will also be influenced by legal or agreed regulations, as well as cost-benefit considerations.

4.2 Procedural requirements

4.2.1 General

Assessment of the objectivity, reliability, validity, sensitivity and diagnosticity of a measurement method can be done by any suitable and scientifically acceptable method. This part of ISO 10075 does not specify one single best way to be followed in this procedure. However, a clear account of the strategy used in assessing the psychometric properties of a measurement method shall be provided in order to allow for an evaluation of the adequacy of the procedure chosen and the results achieved. In fact, the validation procedure will depend on the measurement model assumed, e.g. assuming a probabilistic measurement model will require probabilistic psychometrics, whereas a generalizability approach will require the estimation of variance

components and the calculation of generalizability coefficients (see 4.2.7). There is no single best way for assessing the psychometric properties of an instrument, since this depends in part on the principles underlying the construction of the instrument. However, a generalizability approach allows for effective and efficient procedures to estimate objectivity, reliability, validity, sensitivity and diagnosticity parameters with reasonable effort. In any case, the procedures shall be described in such a way as to enable a replication of the assessment of the psychometric properties.

4.2.2 Objectivity

For documenting the objectivity of a measurement method, evidence shall be provided that the person performing the assessment or measurement has no influence on the scores/observations.This can best be achieved by having more than two intended users of the measurement method performing measurements with this method, and analysing the resulting data by adequate analysis of variance (ANOVA) methods for main or interaction effects of the users of the instrument. The results shall indicate that the corresponding effects do not exist, respectively that the estimated variance components do not deviate significantly from 0. In testing for the presence of such effects, it shall be ensured that any actual existing effect of the person performing the assessment/measurement will not be missed. Therefore, the error probability β should be set to 0,05 and the error probability α should be set at least to 0,20. If effects of the person performing the assessment/measurement cannot be rejected, the measurement method shall not be considered objective.

4.2.3 Reliability

Reliability can be assessed by correlational methods, with both workers and/or situations as the object of measurement, whose relative positioning (within their population or universe) has to be reliably reproduced.

Correlation coefficients do not provide information on differences between data sets at an absolute level. Such information can be obtained using ANOVA methods, which also allow calculation of consistency coefficients, e.g. via intra-class correlation coefficients. Calculating, for example, inter-rater-correlations for a checklist, can lead to high correlations if the raters agree in their relative judgements, and thus to positive estimates of reliability, although these raters may differ considerably in the absolute values in their assessment. Using an ANOVA approach would result in the same correlation but indicate an inter-rater effect.

Homogeneity and/or consistency of a measuring instrument is usually assessed via inter-item- or part-whole correlations, or via correlations between equivalent forms of a measurement method. Stricter models of measurement may require stricter tests, e.g. scalogram analysis. Factor analyses, showing unidimensional measurement, can be used to demonstrate the homogeneity of an instrument.

4.2.4 Validity

The validity of a measurement procedure shall be demonstrated by concurrent validation with a different procedure with demonstrated validity for the intended use.

Where such procedures are not available, validity shall be demonstrated (quasi-) experimentally by varying the conditions/tasks which are postulated to lead to the phenomena/effects to be measured, e.g. varying the amount of work stress which will lead to different degrees of fatigue, if fatigue is the object of measurement. In this case, the intensity of the work stress (e.g. difficulty of task performance) and the duration of the exposition have to be varied, a simple variation of intensity or duration alone is not sufficient for demonstrating validity. A suitable set of tasks (either under field or laboratory conditions) shall be used, together with sufficient times on task in order to produce additive or interactive effects.

NOTE Such sets of suitable tasks may be found in the AGARD[2)] STRES (*Standardized task set for research on environmental stressors*) [1], or CTS (*Criterion Task Set*) [4]. If emotional strain or strain effects are the domain of measurement, groups with different degrees of emotional work stress can be used for demonstrating validity.

2) AGARD: Advisory Group for Aerospace Research and Development.

4.2.5 Sensitivity

The sensitivity of an instrument shall be demonstrated by reproducing experimentally controlled variations in mental workload with regard to intensity, duration and their interaction, e.g. the number of items to be remembered, the number of mental operations to be performed, or the time constraints under which the task has to be performed. Under field conditions, this variation can be produced by selecting conditions that have demonstrated or have been rated by several raters with high inter-rater agreement to produce different degrees of mental work stress.

Grades to be used in the sensitivity study shall be representative of the intended context of use of the measurement instrument.

4.2.6 Diagnosticity

The diagnosticity of an instrument shall be demonstrated by discriminating between different kinds of mental workload, e.g. in the area of the consequences of mental strain between fatigue, monotony, satiation and reduced vigilance. The same applies with regard to demands or specific work stress conditions. Testing diagnosticity can be done experimentally by varying conditions, or under field conditions by selecting such conditions as are known to lead to different kinds of mental work stress, e.g. as shown by task analysis methods or concordant ratings by experts familiar with the task.

The kinds of workload to be used in the diagnosticity study shall be representative of the intended context of use of the meaurement instrument.

4.2.7 Generalizability

If a generalizability approach is used, generalizability shall be demonstrated by using experimental or quasi-experimental approaches, in the laboratory and/or under field conditions, with variations across different conditions, representing variations in the area of interest of measurement (e.g. scores of mental work stress), as well as in conditions known to represent or to contribute to potential relevant measurement error, e.g. time of day, populations, pre-load conditions, experience. Estimating generalizability usually involves estimating variance components from an analysis of variance, depending on the structural model of the score, e.g. its decomposition into the different sources of variation. Thus, the measurement model has to specify in advance what has to be considered error and what target variance. Coefficients of generalizability can then be calculated by relating target variance to the relevant error variance.

4.2.8 Usability

4.2.8.1 General

Measurement methods shall be effective and efficient, and satisfy the needs of the user, i.e. achieve the expected level of precision in the intended measurement area, with minimal effort or capacity requirements, and lead to the desired results.

4.2.8.2 Effectiveness

The effectiveness of the measurement instrument is indicated by its psychometric properties, especially its reliability, validity, sensitivity and diagnosticity.

4.2.8.3 Efficiency

The efficiency of a measurement procedure is a relative concept, depending on the effectiveness and the effort required to perform the measurement, analyse and interpret its results. Measurement procedures with comparable effectiveness should thus be compared according to the effort required to perform a measurement with these procedures. No general quantitative requirements can thus be specified.

4.2.8.4 Satisfaction

In addition to being effective and efficient, a measurement procedure should satisfy the requirements of the users. Complaints about the measurement procedure are an indicator of dissatisfaction, whereas lack of complaints cannot necessarily be interpreted as satisfaction. Thus, it is important to evaluate measurement procedures against the requirements of the users and to document the results with the aim of improving the instrument. However, quantitative requirements cannot be specified, because they would be dependent on the expectations of the (different) users in different contexts.

4.3 Quantitative requirements for measuring instruments

4.3.1 General

In general, those instruments should be chosen which best fulfil the psychometric criteria established in this part of ISO 10075 for a given purpose and intended use.

4.3.2 Requirements concerning objectivity

The results of the measurement shall be free from any significant effects of the person performing, analysing and reporting the assessment. This shall be demonstrated in the validation study, e.g. by analysis of variance, where no main or interaction effects for the assessor shall be found.

4.3.3 Requirements concerning reliability

At least homogeneity or consistency coefficients (e.g. as determined by Cronbach's α for homogeneity or a correlation between parallel forms for consistency) as given in Table 1 shall be achieved for the different levels of measurement.

NOTE Requiring a reliability of 0,7 (as measured by a correlation coefficient) corresponds to a proportion of about 50 % error and 50 % target variance, which is set as the minimum requirement for a reliable measurement procedure.

Inter-rater-reliability requires the same values as consistency reliability.

Stability depends on the time interval between measurements and the intervening events/developments. So stability between measurements in directly adjacent measurement periods should be comparable to the above-mentioned requirements for homogeneity and consistency (except for subjective scaling, where these requirements hold only for parallel forms).

4.3.4 Requirements concerning validity

In concurrent validation (i.e. by two or more independent parallel measurement approaches, e.g. two different psychometric scales) validity coefficients as given in Table 1 shall be reached. These values refer to measurements within the same step of the stress-strain-effects process (see ISO 10075:1991) and within the same kind of measurement technique.

Calculating validity coefficients across steps in the stress-strain-effects process is a form of predictive validity and will usually result in lower validity coefficients than within the same step. However, the same limits for validation apply.

If factorial validity is claimed for a multidimensional instrument, correlations between factors/factor scales of $< 0{,}4$ are required.

If validity is claimed for more than one step in the stress-strain-effects process, it shall be demonstrated for all these steps and separate validity estimates shall be given.

4.3.5 Requirements concerning sensitivity

Measurement methods shall be able to discriminate as much as possible between different degrees of workload in the intended area in the stress-strain-effects process, e.g. different degrees of task difficulty from the CTS [4] or the AGARD STRES [1] battery, and not only between extreme states.

A higher number of levels of the intended area of measurement which can be differentiated by a procedure will result in higher sensitivity (i.e. higher sensitivity will enable a differentiation between more states). Table 1 shows the number of levels to be discriminated at each level of precision.

4.3.6 Requirements concerning diagnosticity

Measurement methods shall be able to discriminate between different sources of work stress and different kinds of effects of mental strain, e.g. different tasks in the AGARD STRES battery. Thus, a measurement procedure shall be specific for the intended area of measurement and not contaminated by other areas. Correlations with other areas shall be as low as possible (e.g. fatigue with satiation). Correlations $> 0{,}40$ indicate substantial overlap and thus lack diagnosticity. Table 1 shows the correlations required at each level of precision.

For demonstrating diagnosticity, the instrument has to be applied in situations with different kinds of workload (e.g. different kinds of stress conditions, different kinds of effects), and it has to be shown that the instrument indicates those aspects of workload only for which it is intended to be used.

A well-known procedure for demonstrating diagnosticity is the multi-trait multi-method approach, where multiple methods of measurements are applied to multiple areas of measurement and the diagnosticity is shown by higher correlations within the same area of measurements than between different areas of measurement.

4.3.7 Requirements concerning generalizability

If a G-theory approach is used, generalizability studies (G-studies) shall be as representative as possible for the intended context of use, i.e. all conditions that might be relevant in the context of measurement should be represented in the study, in order to be able to estimate the relevant error variances and thus the generalizablity of the target score.

Reliability under a G-perspective requires a high generalizabilty for the universal score, taking into account different measurement conditions (e.g. time, items, repetitions). Thus, G-coefficients as given in Table 2 for the universal score shall be achieved for levels 3 to 1 respectively.

Concurrent validation using different methods intended to measure the same object according to a G-perspective, requires the demonstration of high factor loadings of all instruments for the object of measurement (universal and/or composite score) in a multivariate analysis. Construct validity can be demonstrated by high G-coefficients, using a strict experimental approach. G-coefficients as presented in Table 2 shall be achieved for this purpose.

For demonstrating sensitivity, the number of levels given in Table 2 shall be differentiated. This requires high generalizability across the levels to be discriminated. In addition, it has to be demonstrated that all levels can be reliably discriminated (not only the extreme ones, where more than two levels shall be discriminated) by presenting the 95 % confidence intervals for each level.

For demonstrating diagnosticity, the measurement instrument should be applied in various different situations or under experimental conditions differing in the particular kind of workload respectively. G-coefficients as specified in Table 2 are required if the measurement instrument is applied in situations or under experimental conditions involving the kind of workload the instrument is intended for. If the instrument is applied under other conditions, G-coefficients should be definitely lower (e.g. near to 0). Using a multimethod design and a multivariate analysis, high factor loadings on the intended areas of measurement shall be achieved for all instruments intended or known to measure these areas, whereas loadings near to 0 shall be achieved for instruments intended to measure other areas, as well as for the instrument under consideration on areas which it is not intended to measure.

4.3.8 Summary of quantitative requirements

Table 1 gives a summary of the quantitative requirements, using a classical test theory, and Table 2 gives this information for a G-theoretical approach.

Table 1 — Quantitative requirements for measurement procedures for different levels of precision using classical test theoretical approaches

Precision level	Objectivity	Reliability	Validity	Sensitivity	Diagnosticity
1 for accurate measurement purposes	no effect	$\geqslant$ 0,9	$\geqslant$ 0,5	$\geqslant$ 5 levels	$<$ 0,10
2 for screening purposes	no effect	$\geqslant$ 0,8	$\geqslant$ 0,4	$\geqslant$ 3 levels	$<$ 0,20
3 for orienting purposes	no effect	$\geqslant$ 0,7	$\geqslant$ 0,3	$\geqslant$ 2 levels	$<$ 0,40

Table 2 — Quantitative requirements for measurement procedures for different levels of precision using a G-theoretical approach

Precision level	G-coefficients				
	Objectivity	Reliability	Validity	Sensitivity	Diagnosticity
1 for accurate measurement purposes	0	$\geqslant$ 0,9	$\geqslant$ 0,9	$\geqslant$ 5 levels	$\geqslant$ 0,90
2 for screening purposes	0	$\geqslant$ 0,8	$\geqslant$ 0,8	$\geqslant$ 3 levels	$\geqslant$ 0,80
3 for orienting purposes	0	$\geqslant$ 0,7	$\geqslant$ 0,7	$\geqslant$ 2 levels	$\geqslant$ 0,70

Where instruments attaining these criteria are not yet available, they should be developed. Until then, those methods with the highest available psychometric criteria shall be used (see 4.1). In any case, instruments with reliabilities $<$ 0,5 shall not be used at all.

NOTE The requirements in Table 1 or Table 2 do not apply to statutory risk assessments.

4.4 Documentation requirements

4.4.1 Documentation requirements for the development of an instrument to measure mental workload

The documentation concerning the development of an instrument intended to measure mental workload shall contain the following information:

— year of development;

— name and affiliation of the developer;

— intended area of measurement (e.g. mental stress, mental strain, mental fatigue, see ISO 10075:1991);

— precision level of the instrument and criteria concerning reliability, validity, sensitivity and diagnosticity;

— theoretical basis of the measurement approach;

— populations/conditions/universes for which the measurement is intended;

— populations/conditions/universes for which the psychometric criteria have been determined;

— psychometric properties of the instrument, e.g. reliablities, validities, generalizability-coefficients;

— the design of the study for determining the psychometric properties;

— the tasks/conditions (and their variation) which have been used for determining the validity;

— the duration of the work and rest periods (and their sequential order) which have been used for determining the validity;

— requirements/constraints concerning conditions under which the measurement shall be performed;

— requirements/constraints concerning the equipment to be used;

— requirements/constraints concerning the training and qualifications of the person performing the measurement (e.g. psychologist, ergonomist, physiologist, management personnel, occupational health specialists, etc.);

— reference standards, if available, for populations, tasks or universes of conditions;

— procedures/conditions to be observed in performing the measurement;

— procedures/conditions to be observed in analysing the data;

— procedures/conditions to be observed in interpreting the data;

— an estimate of the effort (time, people, equipment) required to perform the measurement;

— response rate of the sample in questionnaire studies;

— description of the development procedure, including results from evaluation studies of the efficiency of the instrument and the satisfaction of the users of the instrument;

— provision of the information required to fill in the checklist given in Annex B.

4.4.2 Documentation requirements for reporting mental-workload measurements and assessments

The measurement report shall contain the following information:

— purpose of the measurement;

— description of the measurement instruments [including a brief description of the theoretical background, area of measurement (e.g. mental stress, mental strain, mental fatigue), precision level];

— name and formal qualification of the person responsible for performing the measurement;

— date and time of measurement;

— number, age, gender and experience of participants or employees used when they are the object of measurement are conditions (description of the test sample);

— description of response rate and representativeness of the responding group(s) in questionnaire studies;

— measures taken to safeguard the objectivity of the measurement procedure;

— measures taken to safeguard anonymity/confidentiality (where applicable);

— measures taken to achieve informed consent;

— completion of the checklist in Annex B;

— description of the measurement procedure;

— description of the workload conditions (workplace, task, etc.);

— constraints concerning the measurement;

— comments from the participants or employees;

— peculiarities during the measurement;

— description of results, if applicable per instrument;

— discussion and interpretation of results and conclusions for further action;

— signature and date.

Annex A
(informative)

Additional information concerning generalizability

Generalizability concerns the reliability with which measurements can be generalized to apply to a population of people and/or to one or more universes of conditions. A universe includes all those conditions for which the measurement is considered relevant (e.g. different times of day, different design variants of a display), and for which those in the workplace design are considered a representative sample.

Typically, the score on an instrument or method is used to make a decision about the quality of workplace design or the necessity for accomplishing corrective measures. Two kinds of decisions are possible, i.e. relative and absolute decisions. Relative decisions presuppose a comparative investigation, for example, of different design layouts. In this case, the decision is based on the relative standing or ranking of design variants. This means that relative interpretations address decisions about how much better a design variant is in relation to others. If no design alternatives exist, that means there is only one solution, and the decision is based on the absolute level of its score. Therefore, absolute interpretations address decisions about how well, for example, a workplace is designed with regard to mental workload, irrespective of other design variants. Generalizability theory provides two reliability-like coefficients, describing the precision of each kind of decision, namely:

a) ρ_{δ}^{2} relative G-coefficient;

b) Φ or ρ_{Δ}^{2} absolute G-coefficient.

The G-coefficients describe the accuracy with which it is possible to generalize from an observed score for a sample of measurements to the universe score.

Generalizability theory exceeds the classical test theory by

a) being able to deal with multiple sources of measurement error,

b) being able to deal with systematic components of measurement error,

c) proposing reliability coefficients adjusted to the intended use of the measurements, and

d) differentiating between relative and absolute decisions.

Furthermore, many of the above-mentioned parameters, such as internal consistency, stability or objectivity, can be determined within one study, if an adequate design is used, including for instance items and occasions of measurement and the persons carrying out the measurement as experimental factors. Likewise, generalizability theory can be applied to investigate the sensitivity and diagnosticity of a measurement instrument, starting from an experimental setting, taking different kinds and levels of workload into consideration.

Annex B
(informative)

Checklist for choosing an instrument

When choosing an instrument, the checklist given in Table B.1 may provide assistance in making an informed decision.

Fill in the required information from the information provided in the manual or publications of the instrument. Circle the relevant requirements and check whether this requirement is fulfilled. Where no quantitative requirements are provided, check for the suitability for the intended purpose of measurement.

Table B.1 — Checklist for choosing an instrument

Item		Subclause of ISO 10075-3	Requirements at			Requirements fulfilled		Not applicable	Comments
			precision level 1	precision level 2	precision level 3	Yes	No		
1	Classical test theory								
1.1	Objectivity	4.3.2	No effect	No effect	No effect				
1.2	Reliability	4.3.3	$\geqslant$ 0,9	$\geqslant$ 0,8	$\geqslant$ 0,7				
1.3	Validity	4.3.4	$\geqslant$ 0,5	$\geqslant$ 0,4	$\geqslant$ 0,3				
1.4	Sensitivity	4.3.5	$\geqslant$ 5 levels	$\geqslant$ 3 levels	$\geqslant$ 2 levels				
1.5	Diagnosticity	4.3.6	$<$ 0,10	$<$ 0,20	$<$ 0,40				
2	Generalizability	4.3.7							
2.1	Objectivity		0	0	0				
2.2	Reliability		$\geqslant$ 0,9	$\geqslant$ 0,8	$\geqslant$ 0,7				
2.3	Validity		$\geqslant$ 0,9	$\geqslant$ 0,8	$\geqslant$ 0,7				
2.4	Sensitivity		$\geqslant$ 5 levels	$\geqslant$ 3 levels	$\geqslant$ 2 levels				
2.5	Diagnosticity		$\geqslant$ 0,90	$\geqslant$ 0,80	$\geqslant$ 0,70				
3	Usability	4.2.8							
3.1	Effectiveness	4.2.8.2							
3.2	Efficiency	4.2.8.3							
3.3	Satisfaction	4.2.8.4							

Bibliography

[1] AGARD, *Human performance assessment methods,* AGARD-AGARDograph-308, Neuilly sur Seine: NATO-AGARD, 1989,

[2] BRENNAN, R. L., *Generalizability theory,* New York: Springer, 2001

[3] CROCKER, L., & ALGINA, J., *Introduction to classical and modern test theory*. Chicago, Ill.: Holt, Rinehart and Winston, 1986

[4] Shingeldecker, C.A., *A task battery for applied human performance assessment research.* AFAMRL Technical Report 84-071; Dayton: Air Force Aerospace Medical Research Laboratory, 1984

[5] ISO 9241-11:1998, *Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11: Guidance on usability*

ICS 13.180

Price based on 15 pages