SAMPLE SIZE AND TEST STANDARDIZATION FOR TASK PERFORMANCE ANALYSIS

Reza Karim, Kambiz Farahmand, North Dakota State University

Abstract

In today's industry, many occupations require manpower for both labor and cognitive resources. Due to rapid technological advancement, people are becoming more dependent on cognitive task performance to make critical decisions. It is critical for many operations to design systems so that the effects of physical stress, however minute, on task performance are considered. In this study, a computer assessment tool was developed to evaluate the effect of low-level physical stress on task performance. The effect of stress was analyzed on overall task performance by the subjects who completed the test with and without any exposure to physical stress. The study focused on how sample size was determined and how test procedures could be standardized for data collection.

Introduction

The evaluation of cognitive task performance is very important in the research and improvement of human-machine interfaces for comfort, satisfaction, efficiency, and safety in the workplace. The need for a standardized way of measuring task performance has been well recognized in today's industry. It is recognized that task efficiency and task quality require standardized work procedures, yet an appropriate measure of human task capacity is still a very challenging topic. It has been examined through various studies in the area of neurology, clinical psychology, and human factors on exactly how human task performance is conducted. In this current study, a standardized task method was developed in order to measure the effect of low-level physical stress on cognitive task performance with a greater degree of accuracy.

An extensive review of the literature indicated that there is a lack of standardization on how to design a test for stress -effect evaluation. This study, then, focused on determining if low-level stress has any effect on task performance. Various authors describe stress in different ways. Lazarus [2], for example, defines stress as a feeling experienced when a person thinks that the social or work demands exceed the personal and social resources the person is able to mobilize. Stress and anxiety are core concepts of psychopathology [3]. A diathesis-stress model assumes that most stress-related complications arise from complex interactions between environmental stressors and biological dispositions that can make an individual collapse. Physical load can cause stress and influence operator performance. In the case of a short–duration, high-intensity physical activity, a decrease in accuracy when performing cognitive tasks was observed, such as in the case of map interpretation while running on a treadmill [4]. Human factors researchers recognize the difficulties in defining the construct of physical stress or fatigue and measuring the effect of fatigue under experimental conditions [5].

This study evaluated the effect of physical stress on various types of tasks in the area of general computation, three-dimensional review, vocabulary, pattern recognition, comparison and arithmetic reasoning. It is critical for many operations to design systems such that the effects of physical stress, however minute, on task performance are considered. The authors focused on standardizing the test protocol to establish a guideline for future research. This study focused mainly on how the number of subjects required for the test was determined.

Background

The assessment of task load and stress and the impact on performing a task are very important when individuals are required to perform a specific type of task. Critical decisions made under stressful conditions result in poor performance which could often be catastrophic. It is critical to determine the effect of stress in the demanding fields of aviation, mining, military, transportation, and other industries involved in engineering and critical-thinking processes. Therefore, it is important to develop a standardized tool that is capable of measuring the effect of stress on task performance and is transferable to various types of industries. Accuracy and response time were utilized to find any effect of physical stress on task performance [6]. Task capacity can be measured from objective and subjective queries. Several studies related to human factors show that self-report (subjective) measures can be useful [7].

Sample size determination is an essential factor to validate any new tool. Statistical power tells us if the results of any test are statistically significant or not. A statistically insignificant result with a high statistical power is explained

as either the research hypothesis not being properly selected or there being less of an effect than predicted. The other approach to determining sample size is to run a pilot test [8]. The results from this study should provide a reasonable estimate of the effect of size. A pilot study is not always feasible and, in such cases, previous experience and theories are used to estimate the effect of size.

In many studies related to cognitive tasks, determinations of sample size are not described. Paas & Adam [9] studied two information processing tasks with sixteen subjects. Eight of the sixteen subjects participated in the test of endurance versus interval protocol physical exertion information processing. The other eight subjects participated in rest versus minimal load protocol exertion information processing. The authors did not discuss the process of selecting the number of the samples and statistical power considered in the test. But, the authors were able to find statistically significant results using the F-test. Aks [10] studied the influence of exercise on visual search with eighteen participants and was able to find statistically significant results using an ANOVA. Joyce [11] conducted a time course effect study of moderate intensity exercise on response execution with ten subjects. The authors found statistically significance results using the F-test. But the authors did not discuss statistical power and how the number of participants was determined for the test.

The tasks performed in any industrial facilities are routine and repetitive in nature. It is important to standardize task performance. The Delphi technique is usually applied to reach a consensus level on a problem where it is difficult to solve the problem experimentally or achieve consensus among the users. One of the key features of the Delphi method is that the participants remain anonymous to ensure that the participants are not influenced by others. The Delphi method is generally conducted through mail service or any other media when the participants cannot meet with the researcher. Also, Delphi allows the flexibility needed for participants to provide feedback at their own pace.

The Delphi method was used for collecting and aggregating information from a group of experts on specific questions and issues related to the subject matter [12]. The Delphi method develops a platform for future knowledge and policy for a specific problem. The results from Delphi studies are widely accepted by the research community because of grassroots involvement. Authors used the Delphi method in different types of research topics. Shah [13] applied the Delphi method to develop a graduate-level lean manufacturing course curriculum. Hasson [14] studied issues in nursing research that included preparation and action steps to be taken by nurses. Any assumption that was considered for developing the tool was challenged when the Delphi mechanism was applied to validate it. The Delphi method helps streamline work flow. Scientific merit questions are an essential part of a Delphi study [15]. The diversified viewpoints help to generate interest among the experts to continue to participate and provide feedback.

To increase efficiency and quality of production, standardized work procedures [16] are required. Many authors developed simulation tools to standardize work procedures [17]. A tool that is built with a standard and widely acceptable method to measure task performance is able to measure task performance capacity with a greater degree of accuracy. As the human brain consists of a complex processing mechanism, task capacity measured using a standard tool is useful in different environmental conditions.

Methodology

In this study, the Delphi method standardized the test procedures to use the task capacity tool. The Delphi method considers views from participants involved in different job functions within the same professional field or closely related professional fields related to the research subject. The research protocol consisted of experimental parameters, variables, procedures, experimental characteristics, number of subjects required, subjects' qualifications, time commitments of subjects, equipment requirement, and a physical exertion protocol. Different mechanisms were introduced in the current study in order to standardize the test protocol. The Delphi method was considered here for evaluating the experimental parameters and variables of the research only.

All of the participants in the Delphi study had science and engineering backgrounds. Since the types of tasks considered were general, no specific branch of scientific background was required for testing the participants. A few samples from the questionnaire include: i) which pair of name is the same, ii) add (+): 76543 and 11111, iii) which picture displays flat piece bent, rolled or both? The seven types of tasks described earlier were considered in order to develop the test in the area of problem solving, memory, and situational awareness of the subjects.

The tasks were classified based on Miller's information task functions [18]. Each question was classified into memory, IQ, and problem-solving type based on the information tasks described by Miller. The reason for classifying these into three performance parameters was to differentiate the effect of stress on each of these measures. A total of twenty-one subjects completed the survey. In the first step, the subjects were asked to comment on the test setup, time allocated for each question, total time taken for the test, and

user friendliness. Based on their feedback, time allocation for some of the questions was increased and the tutorial was improved. In the next step, the subjects were asked to evaluate each of the questions in the test to check the task validity. Finally, redesign of the test based on the feedback was recorded. This methodology standardized the procedure to conduct the task capacity test.

This study was broken into two phases: Phase I and Phase II. Phase I tasks and Phase II tasks were identical. The order of appearance of the questions in both of the tests was random. Phase I was considered as performance without any stress. Phase II was considered as performance after stress. There was one experimental trial for each subject in Phase I. Each experimental trial consisted of thirty tests in a random order. Similarly, Phase II consisted of one experimental trial with thirty tests in a random order. The Phase II test followed immediately after ten minutes of light physical work at the set room air temperature and relative humidity. There was a standard protocol described in the test for biking. The Borg scale was used to rate each participant's stress level. The Borg scale has a range of 6 to 20. The participants were all asked verbally to rate their stress level right after biking. A rating of 10 or lower was considered low-level stress.

This study focused on how to have a balanced experimental design for subsequent statistical analysis. It was desired that the same subjects participate in both experimental phases. However, if subjects dropped out after completing Phase I, they were not replaced by other volunteers during Phase II.

Performance Parameters

The sample size can be determined in numerous ways. The approaches considered for this study are described below.

Approach One:
After the task performance measurement tool was developed, five subjects completed the test at the Phase I level, as shown in Table 1. Based on the test results, a minimum number of participants required for the test was calculated from the sensitivity, power, and statistical analysis of the accuracy level. The sample size was determined from the Operating Characteristics, OC, curve [19], as shown in Figure 1. The OC curve is the plot for type II error. The β error is a function of sample size. For a given value of δ (difference of two means), β error decreases as the sample size increases. The task capacity measured in this study used lower-order cognitive tasks. The tasks selected for the test were considered under a general science category and the subjects who participated had science backgrounds.

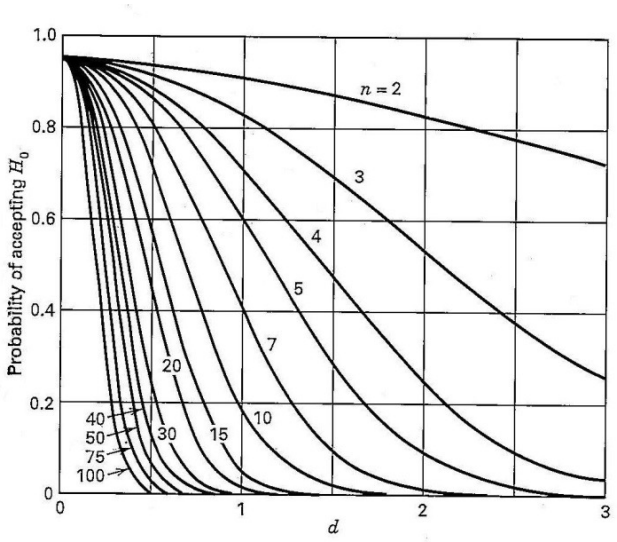

Figure 1. Operating Characteristics Curve [19]

Table 1. Preliminary Data for Sample Size Determination

Overall Correct Answer (%)	Standard Deviation (σ)	Average (μ)	Difference (τ)
85	6.5	84	1
90		84	6
80		84	-4
90		84	6
75		84	-9

From the report of the Army General Classification Test Scores for Civilian Occupation [1], the Binet Intelligence Scale mean IQ values for accountants, engineers, and lawyers are 122 with a standard deviation of 16 and where the minimum value was 96 and the maximum value was 144. The level of difficulty for this test is considered 90% of the maximum score and it can be assumed to be equivalent to the maximum score of 144. The mean score of 122 can be converted to 76.25% of the maximum score with a converted standard deviation of 4.58. Initially in the experiment, it was assumed that task capacity differences between phase I and phase II were not more than 15% of correct results with a standard deviation of less than 5%. The effect size, d, can be calculated using Equation (1).

$$d = \frac{|\mu 1 - \mu 2|}{2\delta} \quad (1)$$

$$= 15/2*4.58 = 1.64$$

From the OC curve (Figure 1) for $\beta = 0.1$ and the d value from Equation (1),

$$n^* = 6.5 \quad (2)$$

where sample size, n, is calculated by

$$n = (n^* + 1)/2 \quad (3)$$
$$= 3.75 \approx 4$$

Based on this approach, the sample size was determined to be four. As the study progressed, it was observed that the mean difference was lower than 15%. Thirty two subjects were considered for this study as the variability between subjects was high and some subjects were expected not to complete both tests.

Approach Two:
A paired t-test was considered in order to determine the sample size using the SAS (Statistical Analysis System) statistical analysis program. Table 2 shows the results from paired-t test runs for mean differences (4), standard deviations (6, 7 and 8), correlations (0.5, 0.6, 0.7 and 0.8), and power considered for the test (0.8).

Table 2. Paired t-Test (Partial Results)

	Computed N Total				
Index	Mean Diff	Std Dev	Corr	Actual Power	N Total
1	4	6	0.5	0.806	20
2	4	6	0.6	0.818	17
3	4	6	0.7	0.807	13
4	4	6	0.8	0.808	10
5	4	7	0.5	0.818	27
6	4	7	0.6	0.828	22
7	4	7	0.7	0.818	17
8	4	7	0.8	0.823	12
9	4	8	0.5	0.852	34
10	4	8	0.6	0.837	28
11	4	8	0.7	0.813	21
12	4	8	0.8	0.818	15

Another SAS program for two sample t-Test for mean differences was used to determine sample size, considering that each subject only completed either Phase I or Phase II of the test. The program was run for a 1:1 ratio and a 2:1 ratio between Phase I and Phase II. The mean differences that were considered were 6, 7, and 8, and standard deviations of 5, 6, 7, and 8. Tables 3 and 4 show the results from the two runs.

Table 3. 2:1 Ratio of Two Sample t-Test (Partial Results)

Computed N Total				
Index	Mean Diff	Std Dev	Actual Power	N Total
1	6	5	0.806	27
2	6	6	0.818	39
3	6	7	0.807	51
4	6	8	0.808	66
5	7	5	0.818	21
6	7	6	0.828	30
7	7	7	0.818	39
8	7	8	0.823	51

Table 4. 1:1 Ratio of Two Sample t-Test (Partial Results)

Computed N Total				
Index	Mean Diff	Std Dev	Actual Power	N Total
1	7	5	0.818	58
2	7	6	0.841	74
3	7	7	0.814	20
4	7	8	0.807	26
5	8	5	0.809	34
6	8	6	0.809	44
7	8	7	0.801	54

Discussion & Results

From the analysis of the literature, it was found that there are no specific criteria for determining the sample size for the Delphi study. Some studies experimented with fifteen subjects with 70% agreement as a consensus level. The Delphi technique was used in this study to determine if the designed questions were covering the seven types of tasks. For each question, a 70% consensus level among the participants was considered acceptable. The Delphi method is a multistage problem-solving method for reducing time and

resources in order to design the questions; existing literature resources were used to determine the types of question used under each task type. The questions which did not receive a 70% consensus level were eliminated from the test. In the third stage, after the participants completed the test, they were asked for any suggestions for improving the test questions. Also participants were asked about their overall experience on the test design. The content validity results are shown in Figure 2. A total of sixteen people responded to this analysis. The initial expectation was to have a content validity rating of 6 or higher in defining 'How important the task is in daily life,' but the plot indicates that some of the question ratings fell below 6. Since the subjects participating in the test were from a wide range of professions, and some participated online, this variability was expected. The plot indicates that only ratings for a few questions fell between 5.6 and 6.

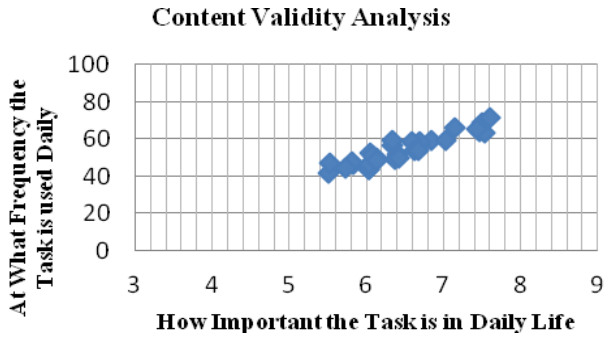

Figure 2. Content Validity Analysis

The test was conducted in a laboratory setting as well as in an online environment. A total of twenty seven subjects completed Phase I and Phase II of the test. Phase I of the test was completed by twelve subjects and Phase II of the test was completed by fifteen subjects.

The ratio of Phase I (including online and lab) and Phase II was approximately 2:1. The differences in mean and standard deviation were 7.25 and 0.47, respectively.

Table 6 shows the results from the sample t-Test SAS program for mean differences of 6 and 7 and standard deviations of 7 and 6. The SAS program considered for the case of 2:1 group ratio at power 0.8 and alpha 0.05. Since both the mean difference and standard deviation calculated from laboratory and online experiments was between 6 and 7, the average N value from Table 6 was considered as total subjects required for the test, which was approximately forty.

Table 5. Phase I Accuracy Analysis

Test Type	Number of Subjects	Mean Accuracy	Standard Deviation
Online Test Average	15	78.4	6.02
Laboratory Test Without Stress (Phase I)	12	78.3	6.6
Laboratory Test Without Stress (Phase II)	15	71.1	6.7

Table 6. SAS Two Sample t-Test

Computed N Total				
Index	Mean Diff	Std Dev	Actual Power	N Total
1	6	7	0.807	51
2	7	6	0.828	30

From the literature review on the effects of physical exertion on task performance, the number of subjects considered varied from ten to thirty two. From the analysis of the SAS results, it was reasonable to consider the total sample size of approximately forty for the study.

The NASA-TLX rating principle was utilized and modified to develop an overall performance chart to be rated by each subject, as shown in Table 7. The chart was required to be completed by the subjects after Phase I and Phase II of the test. The purpose of the subjective rating chart was to estimate the subject's evaluation of the test in terms of mental demand, temporal demand, performance, effort, and frustration level. The analysis of the report, as shown in Table 8, provides an overview for the test structure and the scope of future improvements for the test. The evaluation of the subjective rating indicates that participants were comfortable with how the questions were designed with a very low frustration level and a medium level of effort.

A short survey form, as shown in Table 9, was used to find how the subjects felt about the test design. The form was completed by the subjects after either Phase I or Phase II of the test. The evaluation report indicated that subjects were satisfied with the computer test model in terms of accessibility, navigation, readability, organization, and total time spend on the test.

Table 7. Overall Performance Chart

Title	Endpoints (1-10) scale	Descriptions
Mental Demand	Low / High	How much mental and perceptual activity was required?
Temporal Demand	Low / High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred?
Performance	Poor / Good	How much do you think you were in accomplishing the goals of the task set by the experimenter?
Effort	Low / High	How hard did you have to work (mentally) to accomplish your level of performance?
Frustration Level	Low / High	How stresses versus relaxed did you feel during task?

Table 8. Summary of Subjective Rating on Test

Type	Rating (1-5) frequency	Rating (1-5) %	Rating (6-10) frequency	Rating (6-10) %
Mental Demand	12	37.5	20	62.5
Temporal Demand	12	37.5	20	62.5
Performance	3	9.5	29	90.5
Effort	15	47	17	53
Frustration Level	20	62.5	12	27.5

Table 9. Evaluation of Computer Test

Type	Rating (1-3) Frequency	Rating (4-5) Frequency
Accessibility	3	97
Navigation	0	68
Readability	9	91
Content Organization	6	94
Total Time spent on participating in the test	25	75

Conclusion

The focus of this study was to develop and standardize a task assessment test needed to evaluate individual task capacity and to determine the appropriate sample size to measure the effect that low-level physical stress may have on task performance. The Delphi method was considered as a tool to determine the needs and skills required in any specific work environment. The Delphi technique utilizes combined individual judgment to address any issue related to an incomplete state of knowledge. The consensus was reached above 70% agreement with eighteen subjects who completed the evaluation. The subjects participating in content validity were from a wide range of professions. The consensus level achieved was about 6 and above, in terms of "how important the task is in daily life". Thirty two subjects were considered for this study based on initial test results and statistical analysis. The number of subjects considered satisfied the research objective to determine if stress had any effect on task performance. The developed tool was capable of assessing the effect that low-level physical stress in various types of industries might have on performance. The assessment tool has the capacity to change its settings to incorporate different expertise levels or task types.

References

[1] Trent, W. C. (1985). *Understanding and Selecting Personnel*. Tyler: Gateway Press, Texas.

[2] Lazarus, R. S. (1990). Theory-Based Stress Measurement, *Psychological Inquiry*, 1(1), 3-13.

[3] Kroemer, K. H. E., Kroemer, H. B., & Kroemer-Elbert, K. E. (2003). *Ergonomics: How to Design for Ease and Efficiency.* (2nd ed.). Prentice Hall.

[4] Hancock, S., & McNaughton, L. (1986). Effects of fatigue on ability to process visual information by experienced orienteer. *Perceptual & Motor Skills*, 62, 491-498.

[5] Gawron, V. J., French, J., & Funke, D. An overview of fatigue. In P. A. Hancock & P.A. Desmond (eds.), (2001). *Stress, workload and fatigue* (pp. 581-595). Mahwah, NJ: Lawrence Erlbaum.

[6] Karim, R., & Farahmand, K. (2011). Test Battery for Evaluation of Task Performance and Situational Awareness. *International Journal of Engineering Research and Innovation*, 3(2), 54-60.

[7] Muckler, F. A. (1992). Selecting Performance Measures: Objective versus Subjective Measurement. *Human Factors*, 34 (4), 441-455.

[8] Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral and biomedical

sciences. *Behavioral Research Methods*, 39(2), 175-191.
[9] Paas, F. G., & Adam, J. (1991). Human information processing during physical exercise. *Ergonomic* 34 (11), 1385-1397.
[10] Aks, D. J. (1998). Influence of exercise on visual search: Implications for mediating cognitive mechanisms. *Perceptual and Motor Skills*, 87, 771-783.
[11] Joyce, J., Graydon, J., McMorris, T., & Davranche, K. (2009). The time course effect of moderate intensity exercise on response execution and response inhibition. *Brain and Cognition*, 71, 14-19.
[12] McKenna, H. P. (1994). The Delphi Technique: a worthwhile research approach for nursing? *Journal of Advanced Nursing*, 19, 1221-1225.
[13] Hiral, A. S., & Tillman, T. S. (2008). An International Lean Certification Role Delineation Delphi Study Applied to Develop Graduate-level Curriculum in Lean Manufacturing, *Technology Interface Journal/ IAJC-IJME Conference*, Volume 9(1).
[14] Hasson, F., Keeney, S., & McKenna, H. (2000). Research Guideline for the Delphi Survey technique, *Journal of Advanced Nursing*, 32(4), 1008-1015.
[15] Powell, C. (2000). The Delphi technique: myths and realities, *Journal of Advanced Nursing*, 41(4), 376-382.
[16] Adler, P. S. (January-February, 1993). Time-and-Motion Regained. *Harvard Business Review*, 97-108.
[17] Farahmand, K., Karim, R., Srinivasan, R., Sajjadi, R., & Fisher, L. (2011). Lean Enterprise Principles Applied to Healthcare. *International Journal of Engineering Research and Innovation*, 3(2), 5-13.
[18] Miller, R. B. (1974). A Method for Determining Task Strategies. *American Institutes for Research in the Behavioral Sciences*. (Report no. APHRL-TR-74-26). Silver Spring, MD.
[19] Montgomery, D. C. (2001). *Introduction to Statistical Quality Control*. (4th ed.). Wiley.

Biographies

REZA KARIM completed his Ph.D. in Industrial and Manufacturing Engineering Department at North Dakota State University. He is currently teaching an undergraduate -level lab on Workstation Design and Time Motion Study and conducting research in improving Healthcare facilities design and management. His research interest is in the area of healthcare, simulation, product design, process analysis, lean application and human factor consideration. He has extensive work experience in the field of equipment design, process design, reliability engineering, Computational Fluid Dynamics (CFD) and project management. Mr. Karim may be reached at Reza.Karim@ndsu.edu

KAMBIZ FARAHMAND is currently a Professor at the Industrial and Manufacturing Engineering and Management at North Dakota State University. He is an internationally recognized expert in Productivity Improvement. Dr. Farahmand has over 28 years of experience as an engineer, manager, and educator. He is a registered professional engineer in the state of Texas and North Dakota. Professor Farahmand may be reached at Kambiz.Farahmand@ndsu.edu