NAVAL HEALTH RESEARCH CENTER

PHYSICAL TASK PERFORMANCE:

COMPLEXITY OF THE ABILITY-PERFORMANCE INTERFACE

R. R. Vickers, Jr.

Report No. 95-30

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CALIFORNIA 92186 – 5122

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

Physical Task Performance: Complexity of the Ability-Performance Interface

Ross R. Vickers, Jr.

Human Performance Department
Naval Health Research Center
P. O. Box 85122
San Diego, CA 92186-5122

Report 95-30, supported by the Naval Medical Research and Development Command, Bureau of Medicine and Surgery, under Work Unit 63706N M0096.002-6417.

SUMMARY

Background

U.S. Navy and Marine Corps personnel perform many different physical tasks in their work. If every task must be considered individually, task diversity poses a significant challenge for realistic models to simulate operational performance. In this case, a separate submodel would be needed for each task performed. Prior research suggests that satisfactory models can be achieved by grouping tasks according to the type and magnitude of ability demands imposed.

Objective

The present study assessed the effectiveness of employing a small number of general dimensions to represent a moderately large set of physically demanding Navy tasks and their relationships to physical ability.

Approach

Structural equation models (SEMs) were developed and tested. A correlation matrix for 18 physically demanding tasks and 6 strength measures provided the basis for evaluating and comparing models. The correlation matrix summarized data from 274 male U.S. Navy personnel tested by Robertson and Trent (1985). SEMs were developed separately for physical task performance measures and for physical ability measures. Those models were combined to estimate ability-performance relationships.

Results

A two-dimensional model fit the task performance data better than alternative one- and three-dimensional models. One performance dimension combined carrying and pulling tasks. The other performance dimension was defined by lifting tasks. A general strength dimension was nearly perfectly correlated with the carrying/pulling performance dimension ($r = .962$) and was strongly related to differences on the lifting performance dimension ($r = .742$).

Discussion

Simple models of physical task performance are feasible. In the present instance, physical task performance could be modeled using three major dimensions, one representing strength, two representing different types of tasks. Substituting 2 general dimensions for 18 specific tasks greatly simplifies the problem of modeling physical task performance. Ultimate models of physical task performance are likely to be more complex than this first approximation. For example, the present model does not include endurance components which are important for some tasks.

The results have important applied implications. The development of viable, accurate task performance models is simplified. Studying a small number of appropriately chosen ability-task combinations can provide results that apply to a much wider range of operational tasks. The results also imply that optimal fitness tests can be designed by selecting representative ability measures from the overall ability domain. Such tests may be used as proxies for performance in laboratory simulations and can be part of physical fitness testing for monitoring readiness trends in the Navy and Marine Corps.

Introduction

Simple, general models may provide useful representations of physical abilities and task performance. Physical abilities can be represented by 3 to 6 general dimensions summarizing individual differences on a range of specific physical ability tests (e.g., Hogan, 1991; Myers, Gebhardt, Crump, & Fleishman, 1993). Similar simplicity may suffice to represent the physical task domain. When people perform two or more physically demanding tasks, individual differences in task performance are moderately to strongly positively correlated (Arnold, Rauschenberger, Soubel, & Guion, 1982; Beckett & Hodgdon, 1987; Robertson & Trent, 1985). This correlation pattern suggests the existence of one or more general dimensions underlying differences in performance on the specific tasks. With these observations as a starting point, this paper examines the utility of general dimensions as summary measures to characterize task performance and to describe physical ability-task performance relationships.

The potential value of modeling ability-performance relationships in terms of general dimensions can be illustrated by comparison to more common approaches to modeling physical task performance. Standard practice for predicting physical task performance involves measuring performance on one or more tasks and administering a battery of physical ability measures. Procedures such as stepwise regression then are applied to select a set of ability measures to predict each task. The result is one predictive equation per criterion with different predictors and/or predictor weights for each criterion. The set of regression equations defines the performance prediction model for the task set.

Models comprised of general dimensions could have several advantages over standard practices. Greater parsimony would be achieved in two ways. Fewer causal variables would be invoked (i.e., a few dimensions rather than a large number of specific predictors or tasks). Fewer parameters would be required to express the relationships between the predictors and dependent variables. One practical effect of increased parsimony would be greater precision in estimating parameter values in the predictive equations (Bentler & Mooijaart, 1989).

Increased robustness of the models is a second possible advantage of dimensional models. The estimation of a large number of parameters is implied in the standard approach to performance modeling. The large number of parameters provides a significant opportunity to capitalize on chance relationships to overfit the data when estimating predictive equations. The number of predictor variables is a component of standard shrinkage formulae for multiple correlation coefficients (cf., Schmitt, Coyle, & Rauschenberger, 1977). The more predictors, the greater the expected shrinkage. Thus, all other things equal, an approach which summarizes findings with fewer predictors will yield predictive equations which are closer to true population values for the multiple correlation coefficient and which will cross-validate better than equations derived from an approach with more predictors. Reducing a large number of raw variables to a few dimensions can result in this type of robustness.

Improved variable sampling is a third possible advantage of dimensional models. If diverse physical ability measures are indicators of a few underlying abilities, knowing the structure of those abilities provides a basis for sampling predictors in a given study. Systematic sampling would ensure that all relevant fitness elements were represented in the predictor pool.

Improved variable sampling also might result on the task side of the ability-performance equations. Task performance has not been subjected routinely to the same type of dimensional analysis as abilities, but logical and empirical reasons exist which support the belief that sampling designs can be constructed to increase research efficiency in this domain as well. Common Navy tasks can be grouped into broad conceptual categories such as lifting, carrying, and pulling (e.g., Marcinik, Schibly, Hyde, & Doubt, 1993; Robertson & Trent, 1985). Tasks also tend to be positively correlated when more than one criterion is studied (Arnold et al., 1982; Beckett & Hodgdon, 1987; Robertson & Trent, 1985). Both the logical conceptual structure of the task domain and the empirical evidence of correlations between performance on different tasks are consistent with the inference of one or more general dimensions. Determining the number of dimensions and their composition would provide a basis for sampling tasks within studies.

The advantages of dimensional models can be characterized as "possible" or "potential" gains. These possibilities will be realized only if dimensional models actually can reproduce ability-performance relationships with sufficient precision. Some researchers are skeptical about this possibility (e.g., Robertson & Trent, 1985). Other experts adopt a position implying the utility of general dimensions, but recommend it for the limited purpose of selection screening rather than actual job performance prediction (Vogel, Wright, Patton, Dawson, & Escherback, 1980).

Claims about the feasibility of dimensional models of performance require caution until empirical tests of those claims are available. The present paper provides an initial evaluation by exploring two questions representing critical determinants of the feasibility of the dimensional approach:

a. Can differences in the performance of common Navy physical tasks be summarized adequately by general dimensions?

b. How well do relationships between general dimensions of ability and performance reproduce the observed correlations?

Physical abilities also are represented by general dimensions, but the feasibility of this representation was not regarded as a central research question given substantial prior evidence of general dimensions in this domain (e.g., Baumgartner & Zuidema, 1972; Fleishman, 1964; Hogan, 1991; Jackson, 1971; Jackson & Frankiewicz, 1975; Myers et al., 1993).

METHODS

Data Source

The data analyzed in this paper consisted of a correlation matrix reported by Robertson and Trent (1985). The ability and performance tests and the rationale for their selection are described in detail in that source. For the present purposes, it is important to note that the approach taken in that study began with a survey of experts and job incumbents to identify physically demanding Navy tasks. The tasks were divided into general shipboard tasks which any sailor might be required to

perform (e.g., casualty evacuation, damage control) and tasks specific to particular occupations or ratings (e.g., lifting the canopy on an airplane, loading bombs). The present analysis utilizes data pertaining to 6 strength measures and 18 occupation-specific tasks (Appendix A). The tasks were broadly grouped as carrying, lifting, and pushing/pulling tasks (Robertson & Trent, 1985, p. 14). Examples of each task category were carrying a five-gallon can, raising a canopy on an airplane, and pulling a fuel hose.

Robertson and Trent's (1985) ability measures were chosen to emphasize the dynamic and static strength factors of Fleishman's (1964) strength battery. The present analyses focused on the static strength component because the dynamic strength measures were not included in the test battery for occupational tasks. The measures emphasized arm strength (e.g., armpull strength measured by force on a dynamometer) and lifting (e.g., use of an incremental lift machine (ILM) to establish the maximum weight that could be pressed above the head).

Data analyses reported in this paper employed the correlations between strength and performance measures reported in Appendix E of Robertson and Trent (1985). Although correlation matrices were provided for both males and females, analyses were restricted to the correlation matrix generated by males (N = 274). Robertson and Trent (1985) noted that up to 71% of females had missing data for certain tasks. Exploratory attempts at structural modeling indicated that the resulting correlation matrix was ill-conditioned. Ill-conditioned matrices pose statistical problems for structural modeling (Wothke, 1993). Using the females' correlation matrix in the analyses could have introduced these problems, thereby making the relationship between ability and performance more difficult to specify. These potential problems were avoided by restricting the analysis to the males' correlation matrix.

Structural Equation Models

Structural equation models tested alternative representations of strength and performance, and the relationships between the two domains. Structural models have two components, a measurement model and a substantive model. The measurement model specifies the relationships between observed or measured variables and one or more latent traits. The latent traits represent theoretical constructs invoked to account for the observed pattern of data among indicators for each construct. The substantive model describes relationships between the latent traits. In the present application, for example, the latent traits might represent "strength" and "carrying performance" as exemplified in several strength and task measures, respectively. The substantive model then would deal with the question of how strength was related to performance.

Structural models can be defined at several different levels of specificity. The minimum specification consists of fixing the number of latent traits and defining which measured variables are indicators for each trait. If more information is available, the model can include specification of particular values for model parameters. These parameters include the loadings of measured variables on the latent traits and associations between the latent traits (cf., Bollen, 1989; Joreskog & Sorbom, 1989).

Structural equation models are evaluated by how well the model

accounts for observed covariations between indicator variables. Differences between the observed covariations and the reproduced covariations derived from the model define the fit of the model to the data. These differences can be the result of errors in model specification, parameter estimation, or sampling variability (Browne & Cudeck, 1993). Discrepancies between the model estimates and the observed covariations are summarized by chi-square tests (Bollen, 1989; Joreskog & Sorbom, 1989). However, chi-square evaluations are sensitive to sample size (Marsh, Balla, & McDonald, 1988). Therefore, other measures of fit also are employed. These alternative measures are akin to "variance explained" criteria in commonplace bivariate and multivariate analyses. The Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), a widely-used measure of SEM fit, was the primary model selection criterion in the present analyses. Parsimony adjustments were applied to avoid bias toward the acceptance of more complex models (Mulaik et al., 1989).

Modeling followed the two-stage process recommended by Anderson and Gerbing (1988). First, measurement models were defined for strength and performance. These measurement models consisted of two elements. First, the models specified an hypothesized number of latent trait dimensions for the strength or performance domain as appropriate. Second, the models matched each measured variable to one of the latent trait dimensions and specified that the variable would have a nonzero factor loading on that dimension. Scaling for the factor loadings was established by fixing the variance of the latent trait at 1.000. This choice of scaling procedures permitted the estimation of factor loadings for all of the indicators defining a dimension. The resulting set of factor loadings defined the basic measurement model. As described in the results, this measurement model specification procedure was applied several times in each measurement domain by specifying different numbers of latent traits with different associated patterns of factor loadings.

The second analysis stage examined ability-performance relationships. The measurement models developed in the first stage were fixed elements of a larger SEM in this second stage, i.e., the number of factors, the loadings of indicators on factors, and correlations between the factors within a given domain (e.g., within the performance domain) were fixed at values estimated in the measurement model phase. Strength-performance relationships were determined by examining correlations between the latent traits defined and fixed in the measurement models. These interdomain latent trait correlations were the basis for estimating how strongly performance was related to strength. This sequence of modeling separates the specification of auxiliary measurement models from the estimation of the central substantive relationships (Meehl, 1990).

The computation fit indices for the ability-performance models was not based on the overall chi-square for these models. Instead, these computations focused on the (mis)fit of the model in reproducing only the observed ability-performance associations. This aspect of fit was isolated by decomposing overall model fit indices into separate elements associated with the measurement model and the substantive model. Only the latter source of misfit was relevant to this stage of the model development. The basic procedure for isolating the relevant (mis)fit was:

a. Determine the overall chi-square for a given ability-performance model.

b. Subtract the sum of the chi-squares for the relevant measurement models.

c. Use the resulting difference as the ability-performance chi-square.

Computations for the null chi-square for one model which postulated two strength dimensions and three performance dimensions illustrate the process. The factor loadings necessary to define the base dimensions were obtained from prior analyses undertaken to develop the measurement models. The null model then assumed that all associations between ability and performance dimensions were equal to zero. Fitting this model to the ability-performance covariance matrix produced a chi-square of 1730.80. However, the prior analyses which produced the measurement models indicated that the failure to reproduce correlations between ability measures contributed 10.61 to this total. The imperfect reproduction of covariation between performance measures accounted for an additional 479.07 of the total. The misfit between estimates of the ability-performance correlations and the observed correlations, therefore, was equal to 1241.12 (i.e., 1730.80 - 10.61 - 479.07). Similar computations applied to all models considered.

The null model for ability-performance associations had 108 degrees of freedom. This figure was based on the number of correlations between ability and performance measures (i.e., 18 performance measures combined with 6 strength measures).

Multiple models were considered at each analysis stage. The acceptability of a given model cannot be evaluated solely in terms of the absolute fit between the model and the data (Bollen, 1989). Alternative models must be compared, because several models may be approximately equivalent in representing the data to be modeled. If so, it is misleading to present a single model as if it were the only plausible option. Conversely, a model which fits the data only moderately well in absolute terms sometimes is clearly superior to competing plausible alternatives. In either case, comparison to other competing models provides context for identifying and evaluating the "best" model. The presentation of results, therefore, identifies the best model(s) on the basis of model comparisons rather than relying solely on statistical significance tests or the absolute fit of the model.

Results

The presentation of the results has been structured to mirror the central research questions. First, evidence is presented bearing on whether task performance can be adequately represented by general dimensions. Models for strength measurement then are described as a preface to addressing the second major research question. The question of how strongly ability and performance are related is examined as the third component of the analysis.

Performance Model

Model Specification. Robertson and Trent (1985) classified their tasks as carrying, lifting, or pulling. This classification provided the starting point for the development of three models to represent task performance space. In order of increasing complexity, the models were:

a. One-dimensional Model: Each task performed was assumed to be an indicator for a single general dimension of performance.

b. Two-dimensional Model: The two-dimensional model combined carrying and pulling tasks to define one performance dimension. Lifting tasks defined a second performance dimension. The two performance dimensions were assumed to be positively correlated.

c. Three-Dimensional Model: Carrying, lifting, and pulling tasks defined three distinct dimensions. All pairwise correlations between dimensions were assumed to be positive and greater than zero.

The first and third models were specified a priori; the second model developed from the analyses. The one-dimensional model was the simplest plausible representation of performance differences. The task performance correlation matrix was a positive manifold as would be expected if all of the tasks measured a single underlying construct. Also, ratings of strength requirements for military tasks often produce high correlations for upper and lower body strength (Gebhardt, Jennings, & Fleishman, 1981). This pattern suggests a general dimension of task demands.

The carrying-lifting-pulling model was a conceptual task classification derived from rational analysis (Robertson & Trent, 1985). Tasks in the different categories conceivably could require exertion of different muscle groups or combinations of muscle groups. This model provided a simple representation consistent with other task classifications (e.g., U.S. Department of Labor, 1977).

The two-dimensional performance model was introduced because the three-dimensional model indicated a very high correlation ($r = .961$) between the carrying and pulling latent trait dimensions. The latent trait dimension for lifting was relative moderately related to the latent trait for carrying ($r = .744$) and the latent trait for pulling ($r = .688$). The correlation between lifting and pulling performance was high enough to consider combining the two types of tasks into a single dimension.

Model Evaluation. The three-dimensional model produced the best raw fit to the data (Table 1). The chi-square for the three-dimensional model was significantly less than that for the two-dimensional model (chi-square = 7.38, 2 df, $p < .025$). However, this small difference was not enough to make the three-dimensional model superior by all criteria. The TLI for the two-dimensional model was equal to that for the three-dimensional model. The PTLI favored the two-dimensional model because that model retained more degrees of freedom. The one-dimensional model was the weakest of the three models for reproducing the correlations, but even this model was only slightly poorer than the other two when the TLI and PTLI values were considered. All three models were retained for further analysis to explore the effects of varying the number of dimensions used to represent the task space.

Table 1

Comparison of Performance Measurement Models

Models	df	Chi-square	TLI	PTLI
One-dimensional	135	539.89	.781	.689
Two-dimensional	134	479.07	.812	.711
Three-dimensional	132	471.69	.812	.701

Note: "df" = degrees of freedom; TLI = Tucker-Lewis index; PTLI = parsimony-adjusted TLI. Null model chi-square = 2244.11, 153 df.

Strength Measurement Model

Strength measurement models included:

a. One-dimensional: All strength measures were assumed to be indicators of differences in a single underlying general strength trait.

b. Orthogonal ILM + Arm: All ILM measures were assumed to represent a single underlying dimension. All arm strength measures were assumed to represent a single underlying dimension. The two dimensions were assumed to be independent.

c. Oblique ILM + Arm: The definition of dimensions was the same as that for the two-dimensional orthogonal model. However, the dimensions were assumed to be correlated.

d. 'g' + Arm: All strength measures were assumed to be indicators of differences on a general strength dimension. The arm strength measures also were assumed to be indicators of differences in strength specific to the arms. The dimensions were assumed to be independent.

e. 'g' + ILM: This model was the same as the 'g' + Arms model, except that the second factor was defined by the ILM measures.

The chi-square statistics divided the strength measurement models into three categories based on raw goodness-of-fit to the data (Table 2). The one-dimensional model and the orthogonal ILM + Arm model fit the data comparably. The oblique ILM + Arm model and the 'g' + Arm model fit the data substantially better than either of the first two models, but not as well as the 'g' + ILM model.

Despite obvious differences in raw fit, the results provided no strong basis for choosing one model over the others. The 'g' + ILM model provided the best absolute fit to the data, but utilized more parameters to achieve this fit than did some other models. The effect was that the 'g' + ILM model produced a relatively low parsimony-adjusted fit (i.e., PTLI value). At the other extreme, the one-dimensional model was the

Table 2

Comparison of Alternative Strength Assessment Models

Models	df	Chi-square	TLI	PTLI
One-Dimensional	9	172.40	.784	.470
Orthogonal ILM + Arm	9	164.18	.795	.477
Oblique ILM + Arm	8	32.88	.963	.514
'g' + Arm	6	31.69	.949	.380
'g' + ILM	6	10.61	.991	.396

NOTE: "df" = degrees of freedom; TLI = Tucker-Lewis Index; PTLI = parsimony-adjusted TLI. Null model chi-square = 1730.80, 30 df.

poorest fitting alternative. This model actually had a higher PTLI than the best fitting 'g' + ILM model. The one-dimensional model also was of interest in the context of claims that a general strength dimension is one of three major dimensions of fitness (Hogan, 1991). Given that even the best model was mediocre by at least one criterion and that even the worst model had theoretical merit, all five strength measurement models were carried forward for further examination.

Strength and Performance

Model Specification. The next analysis stage combined the strength and performance measurement models to address the second major research question. The measurement models described above were used to avoid confounding the substantive relationships between strength and performance with measurement model associations (Anderson & Gerbing, 1988). Measurement parameters (i.e., latent trait factor loadings and factor correlations) were fixed at the values estimated in the prior analysis.

The modeling of strength-performance relationships proceeded by adding more parameters representing relationships between strength and task performance. One objective was to estimate associations between the general dimensions of measured strength and task performance. A second objective was to determine how well the limited number of relationships between higher-order dimensions could summarize and account for the bivariate correlations between pairs of strength measures and task performance measures.

Model Comparisons. The model space explored included 15 alternative models produced by combining the 5 ability measurement models with the 3 performance measurement models defined in earlier analyses (Table 3). Given this many models, it is helpful to adopt a systematic evaluation of alternatives. Examination of the results suggested that it would be constructive to compare the models first from the perspective of identifying the best strength model. The subset of models involving the best strength model then is examined to identify the best overall model.

The 'g' + ILM model was consistently superior to other strength models. This point is illustrated by considering the one-dimensional performance model. The 'g' + ILM model fit the data substantially better

Table 3
Performance Prediction Models

Models	df	Chi-square	TLI	PTLI
One-dimensional Performance				
'g' Only	107	558.31	.309	.306
Arm + ILM Orthogonal	106	379.69	.577	.566
Arm + ILM Oblique	106	240.62	.792	.777
'g' + Arm	106	244.11	.787	.772
'g' + ILM	106	216.49	.829	.814
Two-Dimensional Performance				
'g' Only	106	479.88	.420	.412
Arm + ILM Orthogonal	104	305.44	.682	.657
Arm + ILM Oblique	104	163.20	.906	.872
'g' + Arm	104	163.98	.905	.871
'g' + ILM	104	138.75	.945	.910
Three-Dimensional Performance				
'g' Only	105	473.48	.434	.422
Arm + ILM Orthogonal	102	302.39	.683	.645
Arm + ILM Oblique	102	157.84	.912	.861
'g' + Arm	102	158.68	.910	.859
'g' + ILM	102	132.10	.952	.899

NOTE: "df" = degrees of freedom; TLI = Tucker-Lewis Index; PTLI = parsimony-adjusted TLI. Null model chi-square = 1730.80, 30 df.

than the Arm + ILM Oblique model (chi-square difference = 24.13) even though the models had the same degrees of freedom. The chi-square differences between the 'g' + ILM model and the remaining models were large enough to make the fit indices (TLI = .829; PTLI = .814) larger than the corresponding indices for any competing model. The same pattern was obtained for the two-dimensional and three-dimensional models of performance, but the TLI and PTLI values were higher (> .898) for those models.

The two-dimensional model was the best alternative performance model when all criteria were considered. This choice was not as clear cut as the choice of the 'g' + ILM model, because the three-dimensional model was superior to the two-dimensional model by some criteria. The two models were closely comparable in overall fit to the data, and the TLI value for each model was well above the .900 value Bentler and Bonett (1980) suggest as adequate. The virtual equality of fit for the alternative models made it reasonable to treat them as plausible competing alternatives. Since both models were plausible and neither was clearly preferable to the other, parsimony was given substantial weight in selecting a final model. Therefore, the combination of the 'g' + ILM strength model with the two-dimensional performance model was adopted for further analysis. This choice has the added advantage of retaining the two measurement models with the highest PTLI values (Tables 1 and 2).

Physical Ability Correlates of Task Performance

The chosen model provided a simple representation of physical ability and task performance. The 'g' component carried almost all of the predictive power provided by the strength model (Table 4). In fact, the 'g' correlation implies a predictive equation that accounted for 85.9% of the variance in carry/pull performance. The ILM contributed a trivial 0.3% incremental variance explained. The corresponding figures for lifting tasks were 62.9% and 3.5%. Thus, the ability-performance modeling space reduced to a 1 x 2 space for applied modeling purposes.

Table 4
Strength-Performance Latent Trait Correlations

	Performance Latent Trait Dimension	
Strength	Carry/Pull	Lift
'g'	.927	.793
ILM	-.052	.188

Note: The t-value exceeded the recommended critical value for retaining an effect in a structural model (i.e., $t > 2.00$, cf., Joreskog & Sorbom, 1989) for all but the correlation between carry/pull and ILM. The carry/pull and lift dimensions were positively correlated (r = .726).

Final Model

The preceding sections described how well various models fit the data. Figure 1 provides the parameter values for the best fitting model. The parameters include the factor loadings defining the measurement models and the correlations between the latent traits.

DISCUSSION

How well did dimensional models characterize individual differences in the performance of physically demanding Navy tasks? Simple dimensional models did surprisingly well. Individual differences in task performance were adequately represented by two dimensions. This representation was more complex than the unidimensional representation obtained with expert ratings of physical task demands (Gebhardt et al., 1981). At the same time, the model was substantially simpler than treating each task as a separate variable.

Are simple ability-performance models feasible? Such models are very promising. Model simplicity was extended by the finding that almost all of the prediction of task performance differences was provided by a single ability dimension. Although two strength dimensions were identified, only the 'g' dimension defined by the full set of strength measures was important for predicting performance. Thus, the final strength-performance model basically reproduced 108 observed correlations between strength and performance measures with only parameters to describe ability-performance relationships.

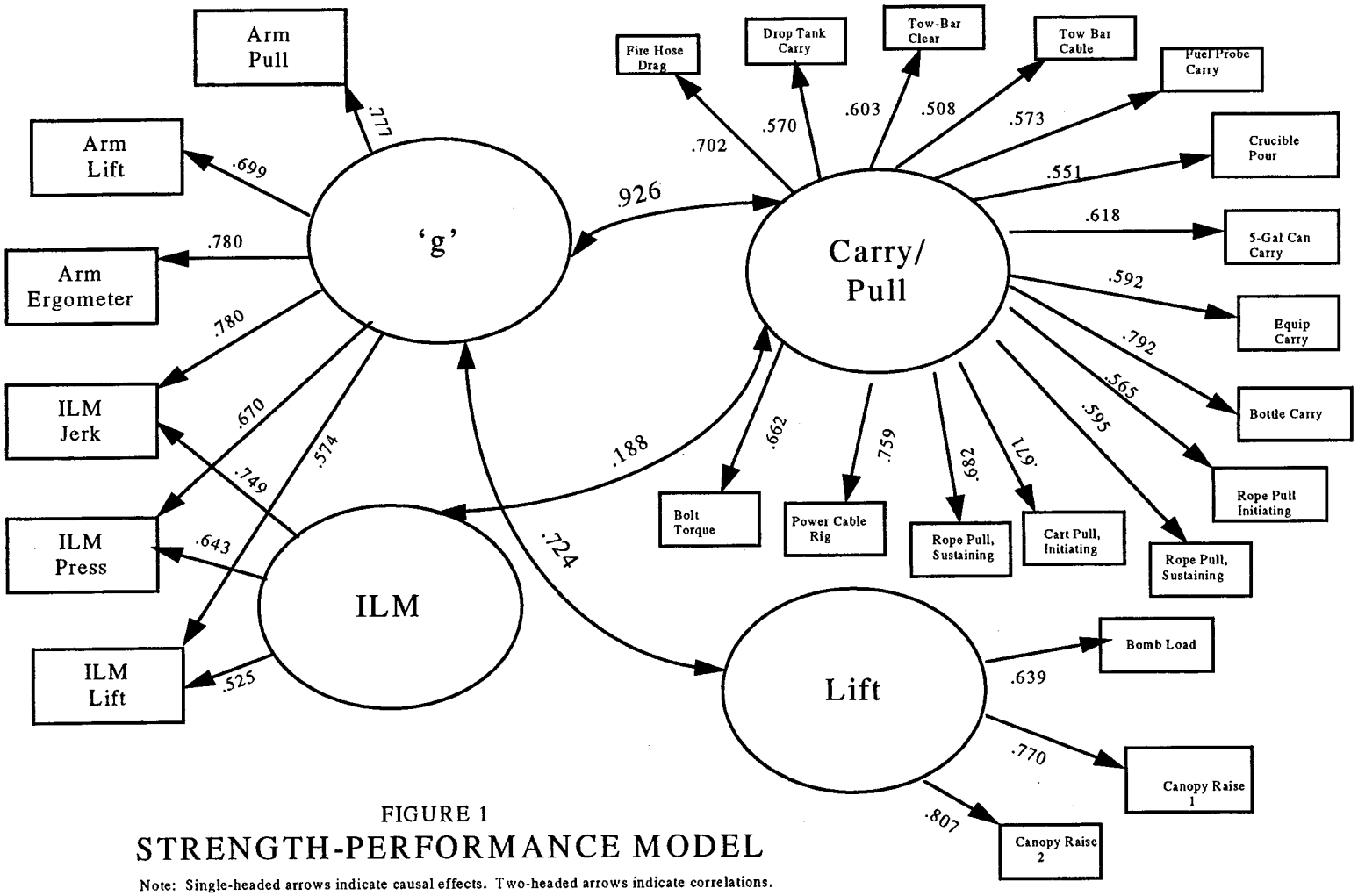

FIGURE 1

STRENGTH-PERFORMANCE MODEL

Note: Single-headed arrows indicate causal effects. Two-headed arrows indicate correlations.

How well did the model represent ability-performance correlations? The fidelity of reproduction was very good even though the substantive model was very simple. This point was supported by the TLI value of .991. This value approached the upper limit of 1.000. The upper limit would be achieved if the residual ability-performance correlations were equal to those expected by chance. One interpretation of the TLI is that the model accounted for almost all of the systematic covariation between the strength measures and performance. Even the PTLI (Mulaik et al., 1989) exceeded the .900 criterion recommended as an adequate fit of the model to the data (Bentler & Bonett, 1980). The model clearly was highly effective in reproducing observed strength-performance correlations.

How do the results compare to other ability-performance findings? This question is difficult to answer, because structural modeling is being applied to this area for the first time. However, there are important qualitative similarities to findings obtained using other methods. Stevenson, Bryant, Greenhorn, Deakin, and Smith (1995) have shown that many strength measures can be reduced to a smaller number of factor scores to predict performance on a single box-lifting task. The factor scores did not predict the criterion as well as the data-level variables used to define the factors. The variance in performance accounted for by the regression shrank from 92% with 32 data-level variables to 75% with 4 factors. This loss of predictive accuracy appears to conflict with the present assertion that a few dimensions provide a satisfactory representation of more extensive sets of specific measures. The apparent inefficiency of the factor-based model may be a statistical artifact. Given the sample size (N = 48), application of Wherry's (1931) formula to estimate the population R^2 (Schmitt et al., 1977) produced a shrunken R^2 = .727 for the factor scores (R^2 = .727) and a shrunken R^2 = .749 for the data-level measures. The four factors, therefore, extracted approximately 97% of the expected population predictive power of the full set of 32 measures. This virtual equivalence suggests that a few factors can substitute for a wider range of data-level variables. The present findings extend this observation by indicating that tasks, too, can be represented by a small number of factors and that knowledge of the relationships between task and ability factors can reproduce the set of correlations which would be the basis for any task-by-task analysis.

The conclusion that a few factors can be substituted for a wider range of specific physical ability measures also is consistent with the results of studies which have not used factor analysis. In this case, the inference is less direct, but still reasonable. The relevant evidence is that studies using individual ability measures to predict specific criteria typically yield simple predictive equations. Only a few predictors are required to extract the predictive power from a large set of ability measures (e.g., Beckett & Hodgdon, 1987). This trend would be expected if a small number of underlying factors determined performance. Selecting one variable as a proxy for the factor would mean that other variables defining the factor would show little incremental variance when added to the equation. Multiple indicators of a single general dimension might show incremental validity if the task depended on specific abilities in addition to the general dimension or if no single indicator was a sufficiently precise proxy for the general dimension. However, the key point would remain. A few predictors representing the relevant major underlying

dimension(s) of ability would be the expected result. Thus, the simplicity of the present model provides a basis for explaining the pattern of some prior results.

What was the 'g' dimension from this study? This question cannot be answered with certainty, because several alternative interpretations of 'g' are plausible. The domain of strength measures can be described in terms of a large number of relatively specific sources of variance (Fleishman, 1964), a few broad general dimensions (Hogan, 1991), or an intermediate number of dimensions (Myers et al., 1993). The sampling of strength assessments in the study which provided the data analyzed here was intended to cover Fleishman's (1964) static strength domain (Robertson & Trent, 1985). However, this sampling strategy still will yield an accurate measure of overall strength as defined by Hogan (1991) if a general strength model is appropriate (Bollen & Lennox, 1991). A sample of strength measures designed specifically to assess multiple dimensions would be needed to eliminate this ambiguity.

Another interpretation of 'g' is suggested by the fact that the pattern of factor loadings emphasized arm strength measures. These measures had the largest loadings and, therefore, were more strongly identified with 'g' than were the ILM tests. Thus, it might be suggested that 'g' measured primarily an upper body strength. This interpretation could be defended on the grounds that some physical assessment schemes include a distinction between upper and lower body strength (e.g., Gebhardt et al., 1981).

At this time, it is suggested that the 'g' dimension modeled here should be interpreted as an index of overall strength. This inference is defensible given the substantive findings. In particular, the strong relationship between 'g' and the carrying/pulling dimension indicated that 'g' was very strongly related to tasks which logically require lower body strength. This relationship would be hard to explain if upper body strength and lower body strength do not define a single general dimension.

The proposed interpretation of the 'g' dimension has significance for applications of the findings. The implications are evident if one considers how the results might be used for job selection and assignment and in fitness monitoring programs. If upper body strength is the key to task performance, measures of this specific element of strength would be preferable to general strength measures for selection and monitoring of physical fitness. Applied fitness assessment batteries would properly be limited to upper body strength measures. If general strength is the key to task performance, applied fitness assessment batteries should be designed to include measures of both upper and lower body strength. This battery composition then would ensure that the assessment reflects general strength rather than an isolated element of strength. The results support the latter possibility over the former.

The concerns just noted arise in the context of applying the results. Either resolution of the ambiguity would leave the study conclusions intact from a modeling perspective. From this perspective, the critical observation is that a single ability dimension was sufficient to predict performance. The validity of this inference does not depend on the composition or meaning of that ability dimension and is not contingent on the theoretical interpretation of differences along the dimension.

The ambiguity of 'g' is one way that research design limited the inferences that can be drawn from the present analyses. Other limitations also may affect the range of legitimate inferences. Task sampling is one possible limitation. The performance tasks were chosen to represent the most physically demanding tasks in Navy jobs. Performance necessarily involved "giving it all you have." Less demanding tasks may utilize more specific muscle combinations and might, therefore, require more complex representation of strength to predict performance accurately. However, the present results are relevant whenever the objective is to characterize strength requirements for the most demanding tasks in a job. As this objective is a common application of performance prediction models, the restriction may not be unduly problematic.

Subject sampling also may have affected the study results. Most subjects were recruits with no experience at the specific job tasks performed in the study. Experience may permit people to develop strategies which depend on their personal strengths and weaknesses. If they exist, strategic differences might reduce the importance of general strength relative to specific abilities. In fact, the administration of the strength and performance assessments included steps to ensure that the same strategy was used by all participants (Appendix A).

What do the findings mean pragmatically? Subject to the limitations noted above, the following points are suggested. First, performance modeling can proceed effectively by studying a few combinations of ability and performance measures. Satisfactory results can be obtained if the specific tasks and ability measures studied are selected to be representative of the higher-order dimensions. Second, the dimensional structure of abilities and performance can be applied to prior research to validate new models. Validation would consist of showing that the associations reported in prior studies can be reproduced from the current model. The predicted association would be the product of the factor loadings for the ability and the task and the correlation between the task and ability dimensions (Joreskog & Sorbom, 1989). Third, simple strength measurement tasks can be useful proxies for job performance measures. This substitution potential is implied by the nearly perfect correlation between the general strength dimension and the performance dimensions. The use of simple, easily standardized tasks constructed to measure specific elements of abilities has obvious advantages over attempting to select a representative task (e.g., Beckett & Hodgdon, 1987) or set of tasks (e.g., Robertson & Trent, 1985) from among the range of options in the Navy.

A fourth potential application of the findings would link research more directly to operational readiness assessments. The potential for substituting strength measurement tasks for job performance in simulations leads naturally to the design of physical fitness tests to monitor the physical component of operational readiness. Fitness tasks selected or designed to cover the job-relevant ability/performance domains can provide direct measures of operational readiness. Information from routine fitness testing then could be used in modeling to define population fitness characteristics. In addition, standard readiness measures could be administered in the field to estimate operational effects on performance (e.g., after a field exercise, during or after operations involving environmental exposures).

Within the limitations of the study, the results imply that the modeling of physical tasks can be simplified and linked directly to force readiness and operational effects assessment. These payoffs derive from the viability of models based on general dimensions of ability and performance. Important details pertaining to the structure of physical abilities and task performance (e.g., when is endurance relevant) remain to be determined. The introduction described several payoffs for modeling performance that would result from filling in the details. Other potential applications were noted in the preceding paragraph. These payoffs make it worthwhile to further evaluate the use of general dimensions for task performance modeling.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103, 411-423.

Arnold, J. D., Rauschenberger, J. M., Soubel, W. G., & Guion, R. M. (1982). Validation and utility of a strength test for selecting steelworkers. Journal of Applied Psychology, 67, 588-604.

Baumgartner, T. A., & Zuidema, M. A. (1972). Factor analysis of physical fitness tests. Research Quarterly, 43, 443-450.

Beckett, M. B., & Hodgdon, J. A. (1987). Lifting and carrying capacities relative to physical fitness measures (Report No. 87-26). San Diego, CA: Naval Health Research Center.

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

Bentler, P. M., & Mooijaart, A. (1989). Choice of a structural model via parsimony: A rationale based on precision. Psychological Bulletin, 106, 315-317.

Bollen, K. A. (1989). Structural equations with latent variables. NY: Wiley.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 110, 305-314.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (eds.), Testing Structural Equations (pp. 136-162). Newbury Park, CA: Sage.

Fleishman, E. A. (1964). The structure and measurement of physical fitness. Englewood Cliffs, NJ: Prentice-Hall.

Gebhardt, D. L., Jennings, M. C., & Fleishman, E. A. (1981). Factors affecting the reliability of physical ability and effort ratings of Navy tasks (Rep. R81-1). Washington, D. C.: Advanced Research Resources Organization.

Hogan, J. C. (1991). Structure of physical performance in occupational tasks. Journal of Applied Psychology, 76, 495-507.

Jackson, A. S. (1971). Factor analysis of selected muscular strength and motor performance tests. Research Quarterly, 42, 164-172.

Jackson, A. S., & Frankiewicz, R. J. (1975). Factorial expressions of muscular strength. Research Quarterly, 46, 206-217.

Joreskog, K. G., & Sorbom, D. (1989). LISREL VII (2nd Ed.). Chicago: SPSS, Inc.

Marcinik, E. J., Schibly, B. A., Hyde, D., & Doubt, T. J. (1993). An analysis of physically demanding tasks performed by U.S. Navy fleet divers (Tech. Rep. No. 93-15). Bethesda, MD: Naval Medical Research Institute.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. Psychological Inquiry, 1, 108-141.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

Myers, D. C., Gebhardt, D. L., Crump, C. E., & Fleishman, E. A. (1993). The dimensions of human physical performance: Factor analysis of strength, stamina, flexibility, and body composition measures. Human Performance, 6, 309-344.

Robertson, D. W., & Trent, T. T. (1985). Documentation of muscularly demanding job tasks and validation of an occupational strength test battery (STB) (Rep. No. 86-1). San Diego, CA: Navy Personnel Research and Development Center.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. (1977). A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. Psychological Bulletin, 84, 751-758.

Stevenson, J., Bryant, T., Greenhorn, D., Deakin, J., & Smith, T. (1995). Development of factor-score-based models to explain and predict maximal box-lifting performance. Ergonomics, 38, 292-302.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

U.S. Department of Labor. (1977). Dictionary of occupational titles (4th

ed.). Washington, DC: U.S. Government Printing Office.

Vogel, J. A., Wright, J. E., Patton, J. P., Dawson, J., & Escherback, M. P. (1980). A system for establishing occupationally-related gender-free physical fitness standards (Tech. Rep. 5). Natick, MA: U.S. Army Research Institute of Environmental Medicine.

Wherry, R. J., Sr. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. Annals of Mathematical Statistics, 2, 440-457.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (eds.), Testing Structural Equations (pp. 256-293). Newbury Park, CA: Sage.

Appendix A
Brief Descriptions of Strength Tests and Simulated Work Tasks

Strength Tests

Arm Pull: Using a push-pull force gauge, participant took handle of gauge in one hand, braced the other against a vertical support, then pulled to determine maximum pull force.

Arm Lift: Using push/pull gauge, subject held lift bar with both hands with forearms horizontal. Subject then exerted as much upward force as possible by flexing at the elbows, legs straight, heels flat, shoulders stable.

Arm Ergometer: Subject turned the wheel on a Monark ergometer as rapidly as possible for 30 sec. with handle arms set at 4 1/2 inches and resistance at 600 KPM.

Incremental Lift Machine, Jerk: Using an Air Force-designed lift machine, subject grasped bar with palms down, knees bent, arms and legs straight, then lifted bar until legs were straight. Initial weight was set based on arm pull score, then increased in 10# increments to maximum weight subject could lift.

Incremental Lift Machine, Press: With bar starting at shoulder level, feet flat, body erect, subject pressed weight to top of head.

Incremental Lift Machine, Elbow: Subject grasped bar on deck with palms up, then stood erect with feet flat and back straight. With bar hanging at knuckle height, subject then raised bar by flexing arms to 90 degrees maintaining posture of feet flat, knees straight, and back erect.

Performance Tasks

Drop-Tank Carry: A gripping device that simulated a tail fin of a drop-tank was attached to a weight of 100#. Using the device as a handle, the weight was 100' in one direction, then 100' back to original position after about a 30-sec rest.

Tow-Bar Run, Clear: An aircraft nose gear two bar with a weight of 62# at the grip point was carried or pulled 300'.

Tow-Bar Run, Cable: Same tow-bar equipment is carried or pulled 300', but must be taken over 1 1/2" pipes simulating aircraft carrier arresting cables.

Fuel Probe Carry: Carry an object with a cylindrical base (12.5" diameter; 2" depth) for 50', rest 30 sec., return to starting point. Weight of 50, 69, 88, 114, 120 pounds selected by subject as heaviest with which he believes he can perform the task.

Crucible Pour: Using handles, slide a simulated crucible 20' along a track walking/stepping sideways. Return to initial position stopping every 2' to rotate the handles 45 degrees to simulate pouring. Weights for load were 99, 130, 153, or 168 pounds (chosen by subject).

5-Gallon Can Carry: Carry 5-gallon can 170' over level surfaces and up and down 2 inclined (not vertical ladders). Load in the can was 0, 35, 45, 60, 75, or 95 pounds with subject choosing heaviest weight he felt he could carry.

Equipment Carry: Carry a weight with a handle to simulate carrying tool or weapons system component. Weight of 70# or 119# was chosen by subject and carried 110' on level surface, and up and down a ladder.

Acetylene Bottle Carry: Gripping device attached to a cart designed to ride on tracks which must be carried/pushed up 7 steps of a ladder. Loads for the cart could be 88, 106, 133, or 150 pounds.

Mark 82 Bomb Loading: Loaded weight bar is lifted first to a mid-point rack on a weight lifting device, then to the top rack. Weights could be 30, 50, 70, 90, 120, 140, 160, or 180 pounds. Weight increased until subject cannot lift next highest value, but can repeat the value just completed.

Canopy Raise, 1-Arm: A canopy-raise simulator is lifted with one hand and a safety strut while standing in fixed inset steps simulating side of plane. Weight of canopy adjusted from 22, 32, 54, 65, 76, 87, 98 pounds to determine greatest weight raised.

Canopy Raise, 2-Arm: Same as 1-arm canopy raise, except both arms may be used while holding safety strut in one hand.

Rope Pull, Initiating Force: 25' rope attached to resistance device set at 160# is pulled 10' as rapidly as possible.

Rope Pull, Sustaining Force: 25' rope attached to resistance device set at 60# is pulled 20' as rapidly as possible.

Cart Pull, Initiating Force: Using handle bar grip attached to same resistance device used in rope pull, pull handle 30' with resistance set at 75#.

Cart Pull, Sustaining Force: Using same equipment but with resistance of 45#, pull handle 100'.

Fuel Hose Drag: Using handle bar grip and resistance set at 105#, pull handle 80'.

Power Cable Rig: Using grip device simulating a 3' diameter, 80# segment of power cable with resistance at 100#, lift an pull cable device 40'.

Bolt Torque: Using a resistance device to assess the torque generated, simulate turning a wrench with one arm braced against upright support.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
	5 Sep 95	Interim Oct 94 - Sep 95

4. TITLE AND SUBTITLE
Physical Task Performance: Complexity of the Ability-Performance Interface

5. FUNDING NUMBERS
Program Element: 63706N
Work Unit Number: M0096.002-6417

6. AUTHOR(S)
Ross R. Vickers, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P. O. Box 85122
San Diego, CA 92186-5122

8. PERFORMING ORGANIZATION
Report No. 95-30

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Medical Research and Development Command
National Naval Medical Center
Building 1, Tower 2
Bethesda, MD 20889-5044

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This paper reports the use of structural equation modeling to predict performance on physically demanding Navy tasks. Simple models are preferable to complex models unless the complex model is substantially more accurate. Typical methods of modeling physical tasks performance can produce complex models even when simple models are more appropriate. One objective of this study was to determine whether task performance could be represented by a small number of dimensions rather than a larger set of individual performance tasks. A two-dimensional model was identified which assigned 15 carrying and pulling tasks to one dimension and three lifting tasks to a second dimension. A second objective was to determine how well general dimensions of fitness predicted differences on the two performance dimensions. A general strength dimension was a very powerful predictor of carrying/pulling (latent trait r = .927) and a moderately powerful predictor of lifting (latent trait r = .793). The final model comprised of factor weights linking the tasks and strength measures to the appropriate dimensions and the correlations between the latent traits effectively summarized the bivariate correlations between individual strength measures and performance on individual tasks. The results indicated that simple representations of abilities and task performance based on general dimensions are a viable method of representing the domains and their interrelations.

14. SUBJECT TERMS			15. NUMBER OF PAGES
physical tasks	strength	job performance	21
physical ability	task structure	military	16. PRICE CODE
physical fitness	structural equation modeling		

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited