NAVAL HEALTH RESEARCH CENTER

GENERALIZABILITY TEST OF A PHYSICAL ABILITY-JOB PERFORMANCE MODEL

R. R. Vickers, Jr.

Report No. 96-16

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

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

Generalizability Test of a Physical Ability-Job Performance Model

Ross R. Vickers, Jr.

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

Report 96-16 was supported by the Navy Medical Research and Development Command, Bureau of Medicine and Surgery, Department of the Navy, under work unit 63706N M0096.002-6417. The views expressed are those of the author and do not reflect the official policy or the position of the Department of the Navy, Department of Defense, or the U.S. Government. Approved for public release; distribution unlimited.

SUMMARY

Problem

Models of U.S. Navy physical task performance can be used to improve job design, physical fitness standards, and wargaming verisimilitude. Generic models of human capabilities, limitations, and performance are needed for simulation and modeling (Under Secretary of Defense [Acquisition and Technology], 1995). Vickers (1995) provided a simple, potentially useful model, but the empirical basis for the model left uncertainty about how broadly it applied to different people and tasks. One concern was that a model based solely on data from males would not generalize to females.

Objective

The present study tested the generalizability of the Vickers (1995) model to females.

Approach

Structural equation models were used to represent the pattern of relationships between physical strength and task performance measures. Data were correlations for females reported by Robertson and Trent (1985).

Results

Repeating Vickers' (1995) procedures with the new data set produced a female performance model with the same formal structure as that obtained in males. Male and female parameter values were very similar, and the male parameter estimates reproduced the female data well enough to conclude that males and females can be represented by a single general performance model. Treating the two samples as replications of this shared general model led to model simplification based on replicated areas of misfit. The final result was a model in which general strength was the sole predictor of general job performance ($\underline{r}$ = .955) for both men and women.

Conclusions

A useful first-approximation generic model for physical performance capabilities of U.S. Navy personnel is provided by a general strength dimension for both men and women. General performance degradation under operational conditions can be estimated by determining how those conditions affect overall strength. This model provides a tool for identifying instances where more refined ability assessments are needed to predict specific critical tasks accurately. The model, therefore, has the potential to guide the development of general fitness criteria and the specification of job-specific criteria where appropriate. The model probably applies only to brief tasks emphasizing strength over endurance, but these tasks may comprise the preponderance of Navy work. The present model provides a starting point for dynamic modeling of many U.S. Navy tasks and for extensions to include tasks requiring endurance as well as strength.

Understanding physical ability-job performance relationships is important for many military personnel selection and assignment policies. Ability requirements also can be used to set targets for physical training programs. Sound human performance models also could improve the performance of wargaming models and simulations (Under Secretary of Defense [Acquisition and Testing], 1995).

Simple models may be adequate for many purposes. A simple two-component ability model has been suggested as a basis for screening and assignment in Army jobs (Vogel, Wright, Patton, Dawson, & Escherback, 1980). A recent study by Vickers (1995) suggests that an even simpler model may be appropriate for the physical tasks comprising many Navy jobs. The present study tested the generalizability of those Navy findings, which were obtained in a sample of males, to a sample of female sailors.

Each Navy job encompasses many tasks. The requirements for developing, validating, and applying human performance models are formidable if each task must be modeled individually. Modeling difficulties decrease substantially if general models of job performance encompassing multiple tasks can be developed and validated. Vickers (1995) presented evidence that a simple model consisting of a general strength component and two task performance components (lifting, and carrying/pulling) provided a useful first approximation for predicting performance in a wide variety of Navy tasks for males. This type of model therefore is applicable to the prediction of performance in the task composites that comprise physically demanding Navy jobs.

Theoretical considerations and empirical evidence make it reasonable that simple general models can effectively account for a wide range of task performance differences. From a theoretical perspective, task performance depends on work capacity, a concept which integrates strength and endurance components of ability to predict the length of time that a specific task can be performed at a given rate (Hill, 1993). The implication is that task performance differences can be predicted from knowledge of the strength and endurance demands of the task and the corresponding abilities of the individual. Vogel et al. (1980) employed a similar perspective, but focused on strength and endurance and individual tasks as the level of analysis for their evaluations of Army jobs. If these basic abilities are combined to represent overall ability to repetitively perform a specific task, the result is work capacity. Work capacity is a useful point of intersection between abilities and task performance because it focuses attention on task performance parameters such as rate and duration of work as determinants of how long the person can perform his or her task. When the issue is the assessment or development of work capacity in an individual, the specific elements of physical ability are the proper frame of reference because they can be measured separately (Hill, 1993) and can be modified independently (Gaesser & Wilson, 1988; Jenkins & Quigley, 1992, 1993; Poole, Ward, & Whipp, 1990). However, the performance of specific tasks depends on the integrated combination of the two elements and may, therefore, be better represented by the work capacity concept.

Empirical support for this perspective is provided by studies which show that much of the variance in job performance can be explained by a few

predictors when a battery of physical ability measures is used to predict a job performance criterion. Examples include Beckett and Hodgdon's (1987) finding that just a few predictors from a large pool of ability measures were needed to predict performance in the simulation of a representative Navy task combining lifting and carrying. Stevenson, Bryant, Greenhorn, Deaking, and Smith's (1995) finding that several general ability factors predicted simulated task performance moderately well compared with the results obtained using all of the individual measures in a large ability battery also supports this view. In this latter case, statistical adjustments for shrinkage must be applied to make the point clearly (cf., Vickers, 1995).

Useful models must replicate and generalize. The present study tested the generalizability of Vickers' (1995) model based on data from males by applying it to data from females. This application of the model represents generalization from one distinct population to another. This test of the model is more stringent than a simple replication because the model can fail to fit the data for several reasons. In a straight replication, the true population relationships being modeled would be the same as those in the initial study. Sampling variability and misspecification of the model based on chance relationships occurring in the first sample would be the reasons for failure to fit the new data (Browne & Cudeck, 1993). Shifting to a new population, women in the present case, adds true population differences as a potential reason for failure of the prior model to fit new data. If the model does fit the new data, three tests have been passed rather than two.

The present study tested the generalizability of the Vickers (1995) model at two levels. First, the initial model development procedures were repeated step-by-step. The objective was to replicate the sequence of decisions that produced the initial model if possible. Replication at this level would imply that the same general type of model was suitable for men and women. Second, specific competing models that were evaluated in the original study were applied to the female data. These comparisons set the parameter values for ability-performance models equal to the values estimated for males. This second step went beyond qualitative replication to a direct quantitative replication of the model. The result of these tests was the development of a revised ability-performance model that was simpler than Vickers' (1995) initial model.

Methods

Data Source

The female correlation matrix from Robertson and Trent (1985, Appendix E, Table E-2) provided the data for this paper. The ability and performance measures and the rationale for their selection are described in detail in that source. For the present purposes, it is important that the study began with a survey of experts and job incumbents to identify physically demanding Navy tasks. The tasks were divided into general shipboard tasks 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 utilized data pertaining to 6 strength measures and 18 job tasks (Appendix A). Robertson and Trent (1985, p. 14) grouped the tasks as carrying, lifting, and pushing/pulling tasks. 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 (1985) chose ability measures to emphasize the dynamic and static strength factors of Fleishman's (1964) strength battery. The present analyses focused on static strength because the dynamic strength measures in the test battery were not related to occupational task performance. The measures emphasized arm strength (e.g., Arm Pull strength measured a dynamometer) and lifting (e.g., using an incremental lift machine [ILM] to measure the maximum weight that could be lifted overhead).

Variable Screening

Task performance was assessed in this study using 15 of the 18 task measures. Vickers (1995) employed all 18 measures, but excluded the female data from the analyses because the matrix of performance task correlations was ill-conditioned. This situation could produce significant problems for structural modeling (Wothen, 1993), so the original work was limited to that data which provided the broadest coverage of the task spectrum and which did not suggest the existence of data limitations which might invalidate the model.

The present study accepted narrower coverage of the task domain to improve the manifest data quality. Tasks involving crucible pouring and carrying an acetylene bottle up a ladder were dropped from the analyses because fewer than one in three of the women studied completed them. Exploratory factor analysis of the correlation matrix for the remaining 16 tasks indicated the correlation matrix was ill-conditioned with one eigenvalue less than zero. The anti-image covariances (AICs) for the "initiating cart pull with a 75-lb load" were equal to zero. This result suggested that this task was the source of the problem. Removing this task left a set of 15 performance measures and eliminated the ill-conditioning problem.

The elimination of three task variables calls attention to two factors that should be considered in the assessment of analysis results. First, even though variables with excessive missing data were excluded, substantial missing data remained for some measures. Robertson and Trent (1985) indicated that bivariate correlations were based on between 141 and 258 cases. Computations for the present analyses assumed a sample size of 150. This figure obviously is not literally correct for all of the correlations analyzed. However, the specification of sample size should not affect the choice of models. Model choice depends on relative χ^2 sizes between competing models. The χ^2 for each model is a multiplicative function of sample size and a fit statistic indicating how closely the model reproduces the observed correlation matrix (Bollen, 1989). Sample size, therefore, is a constant in these comparisons. Changing the sample size would yield a proportional increase in all χ^2s, so their relative sizes would stay the same. In effect, model comparisons depend solely on the fit of the model to the data. Note, however, that this assertion does not mean missing data are unimportant. Inconsistencies in the pattern of

correlations introduced by missing data still may affect model comparisons by influencing the fit for alternative models.

The second point is that the elimination of some task variables might raise doubts about whether the same factors were measured in the male and female data. Intuitively, it may seem reasonable to guess that changing the factor composition (e.g., by changing the set of indicator variables) would change parameter values. If so, the confirmatory analyses would be biased in the direction of poor replication. This logic does not apply. If the model correctly specifies the number of latent traits and which indicators are linked to each trait, the parameter estimates are subject to sampling variability (Browne & Cudeck, 1993), but estimate the proper values even if only a nonrandom subset of indicators is utilized (Bollen & Lennox, 1991). Given a correct model, the elimination of several markers should not affect the replicability of the results.

Model Construction

General Approach. The model construction approach in this study was an elaboration on the methods of Vickers (1995). Those methods emphasized a two-step model development process. This process applied Anderson and Gerbing's (1988) recommendation that the development of measurement models should be separated from the use of those models to estimate relationships between latent traits. This approach separates tests of substantive hypotheses from tests of the auxiliary measurement models (Meehl, 1990) by verifying that the measurement models are adequate prior to using them for hypothesis testing.

The general modeling approach also included two other important elements. The scaling of latent traits was accomplished by fixing the variance of those traits at 1.000. This method of scaling the latent traits permits the estimation of factor loadings for all of the measured (i.e., manifest or observed) variables defining each trait. The second element of the modeling approach was that the covariation between latent traits was treated as defining a set of correlations, not causal relationships.

Strength Measurement Models. Four strength models were considered. A unidimensional model treated all six strength tests as measures of a single general ability dimension. This single dimension was labeled 'g_s' to reflect its general nature ('g') across a number of strength ('s') measures. Each other model involved two dimensions. Two models retained loadings for all 6 strength tests on a general dimension and added loadings that defined a more specific dimension that affected only three tests. A "'g_s'+ ILM" model included loadings for all six strength measures on one dimension and loadings for the three ILM measures on a second dimension. A "'g_s' + Arms" model used the three arm strength indicators to define the methods factor. The "ILM" and "Arms" dimensions were constrained to be orthogonal to the 'g_s' dimension. The remaining two-dimensional model consisted of correlated "ILM" and "Arms" dimensions. Each strength test loaded on one factor.

Performance Measurement Models. Three task performance models were considered. A unidimensional model treated all tasks as indicators of a single underlying task performance dimension (hereafter, 'g_p'). A two-

dimensional model combined carrying and pulling tasks to define one dimension with lifting tasks as a second dimension (cf., Vickers, 1995). A three-dimensional model considered carrying, pulling, and lifting tasks as separate dimensions. Robertson and Trent (1985) provided the classification of tasks into carrying, lifting, and pulling groups based on the actions required of the tasks. Thus, the most complex of the three models was based on logical judgments about the tasks. The less complex models collapsed categories found in this initial set of judgments. The two-dimensional model was the final performance model adopted in Vickers (1995).

Strength-Performance Models. Twelve strength-performance models were produced by combining the four strength models with the three task performance models. Factor loadings and latent trait correlations in these models were fixed at values estimated in the preceding evaluations of strength and performance models. The free parameters in the model, therefore, were limited to the correlations between performance and strength dimensions. The analyses included all 12 to reproduce the procedures used by Vickers (1995). This process permitted a qualitative replication of the initial model construction process by determining whether the original sequence of choices between alternative models could be replicated in a new data set.

Constrained Exploratory Analyses and Confirmatory Analyses. The preceding models were evaluated in both exploratory and confirmatory analyses. Both types of analysis imposed the above constraints regarding the number of factors and which indicator variables loaded on which factors. The constrained exploratory analyses treated the factor loadings and latent trait correlations as free parameters to be estimated from the female correlation matrix. The confirmatory analyses fixed the factor loadings and latent trait correlations at the values estimated previously from the male data.

The constrained exploratory analyses were constrained in the sense that the overall model structure was specified in advance. The constrained exploratory analyses were "exploratory" in the sense that the process of estimating the model parameter values from the data "explored" the data within the constraints imposed by the general model structure. The resulting set of parameter values provided the best fit possible for the data set given the structural constraints of each model. This optimum fit provided a basis for determining whether the original decision sequence underlying the Vickers' (1995) model would replicate and how close the true confirmatory models came to achieving optimum fit within the model constraints.

The true confirmatory models occupy a particularly important place in establishing a scientific basis for performance prediction. These models will be falsified to the extent that the present data are inconsistent with any of their elements, including the specification of the number of factors, the pattern of factor loadings and factor correlations, or the size of the factor loadings and correlations. Given that the models will fit the data only if the available parameter estimates apply to the data, these models are stronger than the exploratory models (Meehl, 1990). The fact that fewer parameter values must be estimated means that the models are more parsimonious (Mulaik et al., 1989).

Model Fit Indicators. Models were evaluated by the Tucker-Lewis index (TLI; Tucker & Lewis, 1973) with parsimony adjustments (Mulaik et al., 1989). This index compares models based on their capacity to explain "excess" covariation, where "excess" is the covariation between indicators that is greater than that expected by chance. The basic statistic is a χ^2 representing the difference between the residual correlations and an identity matrix. The residual correlations differ for different models and are determined by a combination of model misspecification and sampling variability (Browne & Cudeck, 1993). Differences between models should be reflected in the model component of the misfit. The TLI adjusts for sampling variability by subtracting the expected large sample value of the χ^2s from the observed χ^2. Thus, differences in model specification accuracy should be the primary source of TLI differences.

Parsimony adjustments allow for the fact that more complex models generally fit data better than simple models (Mulaik et al., 1989). The adjustment is based on the degrees of freedom used to achieve the fit provided by a particular model. The degrees of freedom used corresponds directly to the number of parameter values actually estimated from the data, so this adjustment is directly linked to the parametric complexity of the model *from an estimation perspective*. Emphasis is added, because models involving the same number of parameters do not necessarily have the same parsimony adjustments. In the present case, a confirmatory model and a constrained exploratory model can have the same number of parameters, but differ in parsimony status because the former does not estimate any parameter values and the latter does. The difference in the adjustment reflects the potential for the exploratory analyses to fit the data by capitalizing on chance in estimating values specific to the particular sample. Parsimony-adjusted TLI values are referred to as "ATLI," while the generic term "goodness-of-fit index" or "GFI" is used when referring to the TLI and ATLI together.

Additional Models. Examination of the previouslY described models provided the basis for conclusions about whether Vickers' (1995) findings replicated. The results of that replication provided an opportunity to refine the initial models by considering areas of misfit between the predefined set of models and the data. Vickers' (1995) initial study stopped with the consideration of the models described above because further elaboration would have been based on post hoc identification of areas of misfit between the models and the data. Post hoc modification involves substantial opportunity to misspecify models by capitalizing on chance (MacCallum, Roznowski, & Necowitz, 1992). Given the addition of a second sample in the present study, it was possible to use replicated misfit between the model and the data as a method of guarding against chance-based modifications. Additional exploratory models therefore were developed with the object of exploring potential simplifications of the model through the addition of methods factors. The specific models are described in the presentation of the results.

Results

Strength Measurement Models

With respect to overall fit to the data, all of the two-dimensional strength measurement models were closely comparable and notably superior to the unidimensional model (Table 1). The raw GFI estimates for the alternative two-dimensional models differed by .03 or less. The GFIs for the unidimensional model all were less than that of the poorest two-dimensional model.

Table 1

Goodness-of-Fit Summary for Strength Models

Model	df	Female χ^2	Male χ^2	Female TLI	Female ATLI	Male TLI
Null	15	454.35				
'g'	9	59.24	62.51	.809	.486	.892
'g' + ILM	6	7.24	22.36	1.007	.604	.983
'g' + Arm	6	6.90	24.84	1.008	.605	.978
ILM + Arm	8	13.45	20.10	.977	.521	.988

Note. The male χ^2 is the value obtained fitting the male model to the female data. "TLI" is the Tucker-Lewis index (Tucker & Lewis, 1973) and "ATLI" is the adjusted Tucker-Lewis index. Only the TLI is given for males because the ATLI for these models is identical to the TLI. This identity occurs because these confirmatory models do not estimate any parameters, so the adjustment factor is 1.00 (cf., Mulaik et al., 1989).

The decision sequence described by Vickers (1995) in his analysis of male data replicated well in the female data. The simple 'g_s' model was the weakest alternative, followed by the ILM + Arm model. The only difference was that the female data produced slightly better fit for the 'g_s' + Arm model than for the 'g_s' + ILM model. The order was reversed in the male data, but the difference was slight in these data as well. In light of the inconsistency in identifying the best model, it is noteworthy that the 'g_s' + ILM model actually generalized better than the 'g_s' + Arm model (TLI = .983 vs. TLI = .978).

The male model generalized well to females. The absolute fit of the male model was consistently lower than that for the female model. Male model GFI values ranged from .978 to .988 for the two-dimensional models. These values were substantially higher than the parsimony-adjusted values for the female models and were nearly equal to the raw GFI values for the female models.

Table 2

Goodness-of-Fit Summary for Task Performance Models

Model	df	Female χ^2	Male χ^2	Female TLI	Female ATLI	Male TLI
Null	105	1351.22				
1-dimensional	90	595.33	687.65	.527	.452	.533
2-dimensional	89	555.35	624.66	.559	.473	.566
3-dimensional	87	533.09	646.23	.568	.471	.583

Note. The male χ^2 is the value obtained fitting the male model to the female data. "TLI" is the Tucker-Lewis index (Tucker & Lewis, 1973) and "ATLI" is the adjusted Tucker-Lewis index. Only the TLI is given for males because the ATLI for these models is identical to the TLI. This identity occurs because these confirmatory models do not estimate any parameters, so the adjustment factor is 1.00 (cf., Mulaik et al., 1989).

Performance Measurement Models

Task performance models were less clearly differentiated than were the ability models (Table 2). The range of GFI values was only .041 for the raw GFI and .021 for the parsimony-adjusted GFI. The lower GFIs for these models compared to those for the ability models indicated less explanatory power for the performance models.

The results again replicated the sequence of model choices reported by Vickers (1995) for male data. The 2-dimensional model had the best parsimony-adjusted TLI. The difference was slight, particularly compared with the 3-dimensional model, but the ordering was consistent.

Generalization tests led to a different conclusion than that arrived at by within-sample analyses. The confirmatory generalization analyses indicated that the 3-dimensional model was superior to the 2-dimensional model. The TLI difference between the 2- and 3-dimensional models (.017) was roughly half the difference between the TLIs for the 1- and 2-dimensional models (.033).

Strength-Performance Models

The ability-performance models favored simplicity over complexity on the performance side of the strength-performance equation (Table 3). The performance measurement models suggested that two or three performance dimensions were appropriate depending on the selection criterion used. However, the unidimensional model was the best choice for reproducing the strength-performance relationships. For example, if one considers the 'g_s' + ILM model, the ATLI for the 1-dimensional performance model was .445 compared to .412 for the 2-dimensional performance model and .398 for the 3-dimensional model. The corresponding trend for the confirmatory

Table 3

Goodness-of-Fit Summary for Ability-Performance Models

Model	Female df	Female χ^2	Male χ^2	Female TLI	Female ATLI	Male TLI
1-Dim						
'g'	89	298.22	301.40	.296	.293	.297
'g' + ILM	88	248.08	254.44	.455	.445	.453
'g' + Arm	88	259.35	249.28	.417	.408	.470
ILM + Arm	88	248.04	252.23	.455	.445	.460
2-dim						
'g'	88	299.01	312.91	.282	.276	.258
'g' + ILM	86	249.24	269.32	.432	.412	.403
'g' + Arm	86	260.15	263.88	.394	.376	.422
ILM + Arm	86	249.30	258.96	.431	.412	.438
3-dim						
'g'	87	298.02	314.52	.274	.265	.253
'g' + ILM	84	244.85	271.47	.427	.398	.396
'g' + Arm	84	254.63	266.67	.392	.366	.412
ILM + Arm	84	245.44	268.81	.425	.396	.405
Null	90	390.56				

Note. "Female" refers to results obtained when the model parameters were estimated from the female correlation matrix. "Male" refers to results obtained when the model parameter values originally computed from the male correlation matrix were applied to the female correlation matrix. ATLI is not given for males because TLI and ATLI are identical when a completely constrained model is fitted (i.e., when the analysis uses zero degrees of freedom).

analyses is shown in Male TLI values of .453, .403, and .396, respectively. This pattern was repeated for each of the strength models. Thus, the unidimensional performance model was consistently preferred as a representation of the strength-performance relationships.

The ability-performance results also gave reason to consider all of the two-dimensional ability models. Within the constrained exploratory analysis of the female data, the 'g_s' + ILM model and the ILM + Arm model produced virtually identical raw and parsimony-adjusted fit values. However, the confirmatory application of the male 'g_s' + Arm model generalized slightly more strongly for the 1- and 3-dimensional models. This fact is somewhat offset by the finding that the 'g_s' + Arm model clearly provided poorer absolute fit to the data in the present sample than did either of the competing alternative models.

Male results generalized well to females. The GFI values for these models generally were never more than .031 less than the corresponding ***raw*** GFI for the female data and actually were larger than the raw GFI for 6 of 12 models. The male model TLI was higher than the ATLI for the corresponding constrained female model in 8 of 12 comparisons.

Effects of Eliminating Unnecessary Model Components

The preceding analyses assumed that all correlations between strength and performance latent traits differed from zero. Examination of the estimated correlations indicated that this assumption might be incorrect in some instances. For example, the relationship between the ILM strength dimension and the lifting performance dimension was r = .043. In other models, all parameters were substantially different than zero. Including parameters with true values near zero in a model consumes a degree of freedom without improving the fit of the model. Eliminating the zero effects, therefore, would restore a degree of freedom with little effect on the GFI values. Given that the different models were so closely comparable in the initial analyses, it was desirable to determine whether the seemingly minor effects on degrees of freedom had influenced the comparisons.

Eliminating ILM Effects. The impact of removing parameters linking ILM to performance was evaluated first by eliminating the relationship to the lifting performance dimension. This change increased the misfit of the model only slightly (χ^2 increase = 0.22, 1 df, $p < .640$). Eliminating the correlation of ILM to the carrying/pulling dimension produced a significant GFI change (χ^2 increase = 4.23, 1 df, $p < .040$), but considering the two omissions together would yield a nonsignificant change (χ^2 increase = 4.45, 2 df, $p < .109$). The ATLI increased from .412 for the original model to .426 after eliminating the first correlation, then to .427 after eliminating the second correlation. These latter two figures were higher than the corresponding GFIs for 'g_s' + Arm (.376) and Arm + ILM (.412) models. Those GFIs could not be increased by eliminating latent trait correlations, because neither model included any associations which were close to zero. Thus, with the appropriate elimination of near-zero parameters, the 'g_s' + ILM model clearly was the preferred option.

Examination of the model that combined the 'g_s' + ILM ability model with the univariate performance model suggested another potential modification. The ILM effect on performance could be dropped. This relationship was slight ($r = .131$), but it did pass Joreskog and Sorbom's (1989) t-value criterion for model parameters ($t = 2.06$). Fixing the ILM-performance relationship at .00 increased the χ^2 for the overall model (χ^2 increase = 4.33, $p < .038$). The ATLI (.445) was unchanged compared with the model with both effects included (.445).

ILM effects could be removed from the models with little or no loss of accuracy. These elements therefore were dropped from the model.

Eliminating the Lifting Dimension of Performance. Additional models were considered to account for the better GFI values obtained when modeling performance as a unidimensional construct rather than a two-dimensional construct. The fact that the two-dimensional representation of performance

Table 4

Ability-Performance Associations in Correlated Error Models

		r_{gg}	r_{gI}	TLI	ATLI
Canopy					
'g'	253.44	.925		.447	.442
'g + ILM'	249.06	.891	.132	.452	.442
Canopy & Tow-bar					
'g'	250.03	.932		.473	.468
'g + ILM'	245.53	.898	.137	.448	.438

Note. "r_{gg}" is the correlation between the general latent traits for performance and strength. "r_{gI}" is the correlation between the general latent trait for performance and the ILM latent trait for strength.

included one dimension with only three indicator variables, two of which were variations on a single task (i.e., a canopy raise), suggested a resolution to this point. This lifting "dimension" might really be a narrow construct specific to raising airplane canopies.

The redefinition of "lifting" was tested by modeling performance as a 'g_p' dimension with a correlated error term for the two canopy-raising tasks. This representation fit the data better than the 2-dimensional performance model (537.42, 89 df, vs. 555.35, 89 df). The degrees of freedom were the same for the two models because the introduction of the correlated error was accompanied by the removal of a correlation between the carrying/pulling and lifting dimensions. This revised model fit the data substantially better than the original unidimensional performance model (χ^2 decrease = 57.91, 1 df, $p < .001$) and almost as well as the original 3-dimensional performance model (χ^2 = 533.09, 87 df; χ^2 difference = 3.67, 2 df, $p < .160$). The ATLI for the correlated error model (.488) was better than those for the initial unidimensional model (.452), the 2-dimensional model (.471), and the 3-dimensional model (.473).

Reducing the lifting dimension to a correlated error term did not adversely affect the reproduction of strength-performance relationships. The first line of Table 4 indicates results obtained assuming that performance was related only to the 'g_s' of the reduced model. The second line indicates the results obtained when ability was related to both 'g_s' and ILM. The performance-ILM correlation was small (r = .132), but it was large enough to satisfy Joreskog and Sorbom's (1989) criterion of a t-value greater than 2.00 (t = 2.07) and larger than Cohen's (1969) recommendation that effect sizes greater than .10 be retained as potential elements of predictive models. The change in fit was statistically significant (χ^2 reduction = 4.38, 1 df, $p < .037$).

Two facts weighed against adopting the models with both the 'g' and ILM dimensions despite the marginally significant effects and improvements

in fit noted above. First, the GFI values were higher when only the 'g' ability correlation was involved (Table 4). Second, the 'g_s' only model actually predicted performance better than the 'g_s' + ILM models. When only the canopy-raise correlated error was included, the latent trait r = .925 compared to a multiple correlation coefficient of .901 when the 'g_s' and ILM predictors were used in combination.

Tow-Bar Error. The utility of treating the two canopy raise measures as containing correlated error directed attention to other similar sources of model improvement. Because post hoc model modifications are sensitive to chance covariations between indicators (MacCallum et al., 1992), the search for additional modifications was limited to instances in which methods variance might be a significant source of covariation, i.e., when two very similar measures were involved. The tow-bar carry tasks differed only in that one was over a clear deck while the other involved stepping over cables. Adding a correlated error for these two tasks to the preceding model with just the correlated error for the canopy tasks significantly reduced the model chi-square (χ^2 reduction = 38.51, 1 df, p < .001). Correlations between performance and ability were similar to those seen in prior models. The correlation between 'g_p' and ILM strength was small (r = .137), but met Joreskog & Sorbom's criterion for an acceptable effect (t = 2.13) and produced a statistically significant decrease in the overall model χ^2 (4.50, p < .034).

Adding the tow-bar correlated error to the model produced GFIs consistent with eliminating the ILM association. Fixing this relationship at r = .00 produced the highest ATLI of any strength-performance model (.468 vs. .445 for several previously considered models). Again, the 'g_s' only correlation of r = .932 still was larger than the combined multiple correlation coefficient of .908 obtained with the 'g_s' + ILM model.

Further Modifications. Residuals from the model with correlated errors for the canopy raise and tow-bar carry were examined to determine whether further model modifications were appropriate. The correlated error model was applied separately to the Robertson and Trent (1985) correlation matrices for men and for women. This approach made it possible to identify misfit between the model and the data in both data sets. Replication of misfit across the data sets was used as the justification for any additional modifications. Replication criterion should help minimize the effects of chance in these post hoc modifications.

Unlike the preceding analyses, which employed fixed measurement models, the correlated error models were estimated with all factor loadings, latent trait correlations, and error correlations freely estimated. This approach optimized parameter values within what was not believed to be a reasonable working model of strength and performance. The approach also provided directly comparable male and female analyses without having to repeat all of the previous analyses described in this paper for males using just the 21 variables considered here.

The determination of whether the data indicated a need for further modification proceeded in two phases. The first phase examined the overall pattern of residuals for evidence that factors other than chance were at work. The Kolmogorov-Smirnov (K-S) test (Siegel, 1956) was applied to

determine whether the observed distribution of residuals was normal with a mean of 0.00 and a standard deviation of 1.00. This distribution represented what would be expected if the set of residuals demonstrated only chance deviation from the observed correlations. If the residuals did not conform to the hypothetical distribution for either the male data set or the female data set, two analyses were conducted to determine whether the male and female data sets showed a general tendency toward replicable residuals. One analysis computed the correlation between the standardized residuals for the two data sets. A second analysis recoded the standardized residuals into nominally significant negative correlations ($z < -1.96$), nonsignficant residual correlations ($-1.96 < z < 1.96$), and nominally significant positive residual correlations ($z > 1.96$). Cohen's (1960) κ was used to describe the association between the trichotomies. Two methods were used because they might have differential sensitivity to what might be infrequent events (i.e., replicated residuals).

The second phase of the residuals evaluation examined individual residuals. Specific pairs of variables for which both women and men produced large (i.e., > 1.96 absolute) standardized residuals were identified. Residuals meeting this criterion were candidates for inclusion in the revised model.

Ability-performance residuals were examined first. These residual distributions did not conform to the hypothetical distribution for the male data (K-S $z = 1.88$, $p < .003$), but the hypothetical distribution could not be rejected for women (K-S $z = 0.67$, $p < .766$). A second test which removed the assumptions about the mean and standard deviation of the distribution indicated that the residuals were normally distributed for both men (K-S $z = 1.02$, $p < .250$) and women (K-S $z = .57$, $p < .902$). The rejection of the null hypothesis in the initial test occurred because the variance for males was substantially higher than the expected 1.00 ($SD = 2.39$), but was near 1.00 for women ($SD = 1.23$). The mean residual was quite less than .05 (absolute) for both data sets.

Although the distributions did not conform to expectations for chance misfit between the model and the data, the cumulative evidence provided little or no evidence of ***replicated*** deviations from the model. The Pearson product-moment correlation between standardized scores was $r = .08$ ($n = 90$, $p < .467$). There was no reliable identification of "significant" z-scores ($\kappa = -.05$). Furthermore, only one residual was as large as $z = 1.96$ (absolute) in both samples. The Arm Pull and Bolt Torque tests produced $z = 9.51$ for males and $z = 4.90$ for females.

A correlated error for Arm Pull and Bolt Torque was added to the model despite the lack of an overall pattern of replicable discrepancies. The primary reason for this decision was that the rationale for the inclusion of correlated error terms for the canopy raise and tow-bar residuals correlations applied as well to this pair of variables. Descriptions of the Arm Pull and Bolt Torque measures indicate that these two tests were virtually identical (Robertson & Trent, 1985; see also Appendix A of this report) even though one variable was classified as an ability measure and the other as a task performance measure. The presence of large residuals for both men and women was consistent with viewing this pair of variables as involving task-specific covariance. Adding the

correlated error term to the model reduced the χ^2 by 21.96 for females and 94.32 for males. TLI values increased from .674 to .684 for women and from .682 to .740 for men. The ATLI values increased modestly for women from .587 to .593, but substantially for men from .594 to .641.

Attention then was directed to the performance measures. The initial test for the hypothetical residual distribution indicated a significant discrepancy for men (K-S z = 1.89, p < .002) and women (K-S z = 1.54, p < .018). Here again, tests for normality without the constraints on the mean or standard deviation of the distribution indicated that the residuals were normally distributed for the male data (K-S z = .71, p < .689) and females, (K-S z = .90, p < .400). The average residual was approximately zero for both the male and female data, but both standard deviations clearly were greater than 1.00 for both males (SD = 1.91) and females (SD = 1.85).

There was a detectable trend for task pairs with large residuals in females to produce similar residuals in males. The Pearson product moment correlation was moderate in magnitude (r = .31, n = 103, p < .002), but the trichotomy analysis showed only a weak trend (κ = .06).

The Pearson product-moment correlation gave reason to examine the residuals to determine which ones were extreme in both the female data set and the male data set. The residuals that were large in both samples included: (a) Acetylene Bottle Carry/Rope Pull 160 (male z = 6.58; female z = -3.57); (b) Acetylene Bottle Carry/Drop Tank Carry (male z = 5.29; female z = 2.38); (c) Bomb Load/Canopy Raise 2 (male z = 3.87; female z = 2.32); (d) Rope Pull 160/Power Cable Drag (male z = 2.06; female z = 5.28); and (e) Fire Hose Drag/Rope Pull 60 (male z = 3.13; female z = 3.11). These z-values changed only slightly after the addition of the correlated error for Arm Pull and Bolt Torque.

Addition of the five correlated errors that met the replication criterion produced a substantial change in the fit of the model to the data (females, χ^2 reduction = 56.94; males, χ^2 reduction = 98.59). TLI values increased from .684 to .707 for women and from .740 to .795 for men. ATLI values increased from .593 to .596 and from .641 to .670 for women and men, respectively.

Model modification stopped at this point. The model now included eight correlated residuals (one ability-performance; seven performance). The remaining residuals conformed to the hypothetical distribution for the female data (K-S z = 1.02, p < .250), but not for male data (K-S z = 2.26, p < .001). As in prior analyses, the residuals were normally distributed for males (K-S z = .72, p < .672) and females (K-S z = .79, p < .561), but males produced a much larger variance than expected under the null hypothesis. There was virtually no evidence that the two data sets produced replicable correlated errors (r = .03, p < .664). When trichotomized, κ was -.04. Three residuals did produce large z-values in both data sets, but the absolute magnitudes of those residuals was small. Only one residual was as large as .10 in both data sets. This was less than the 3.1 residuals expected by chance given that 10.5% of the female residuals and 14.3% of the male residuals exceeded .10. Thus, although it

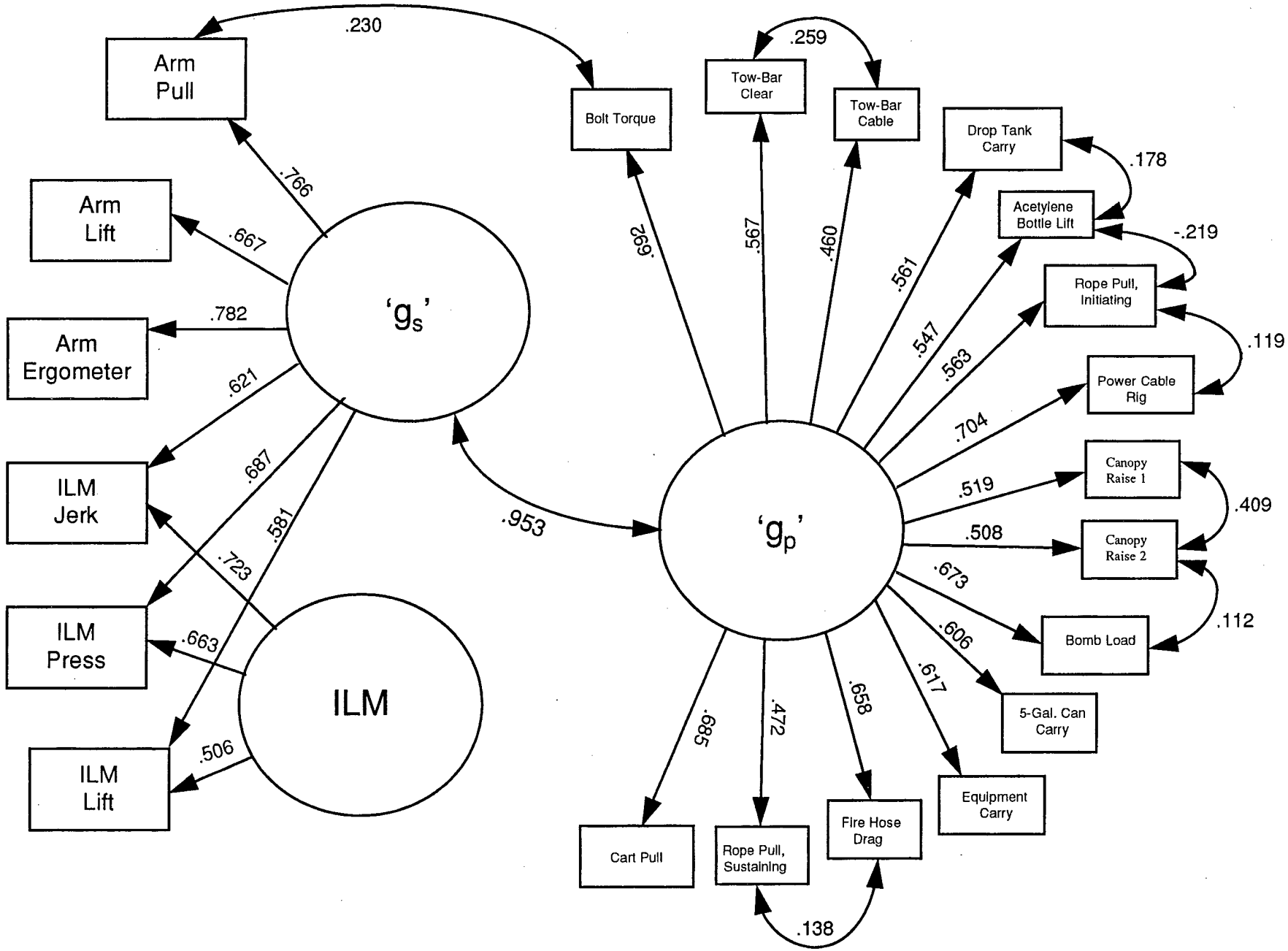

Figure 1. Strength-performance model. Single-headed arrows indicate causal effects of latent traits. Two-headed arrows indicate correlations between latent traits or residuals.

is unlikely that all of the residuals were truly equal to zero, it was reasonable to conclude that any that were omitted from the model were small.

The final model yielded by the data analyses is shown in Figure 1. The parameter values given in the figure were derived by analyzing the male and female data matrices simultaneously as a two-group model. This analysis constrained the parameter values to be equal for males and females. The TLI for the model was .738 compared to .748 when all parameters were estimated freely for the male and female correlation matrices. The difference in TLI values was quite small given that estimating separate models for males and females required 54 more degrees of freedom. This aspect of the model evaluation is clearly reflected in the much larger ATLI for the constrained model (.717) compared to the gender-specific models (.630).

Three points in Figure 1 are particularly noteworthy. The first is that the correlation between 'g_s' and 'g_p' approaches the maximum possible value of 1.00. The second is that correlated residuals were confined largely to the performance domain. Only 1 of 90 strength-performance residuals was included in the model, and this situation arguably was a case of placing the same task in both the strength and performance domains. The third point is that three of five correlated residuals added at the last step in model development were estimated to be less than .14 in magnitude when the full model was estimated. This fact and the analysis of the residuals from this model suggests that the model accounted for all sizable associations between variables.

Additional analyses explored the potential value of further refining the model by permitting selected parameters to differ for men and women. These analyses indicated that very little would be gained by such an approach. Appendix B shows the parameter values obtained by estimating the model separately for male and female data matrices and describes a search for alternative models with different male and female values for selected parameters.

Discussion

This study demonstrated that Vickers' (1995) simple physical ability-job performance model generalized across genders. Direct application of the prior model with measurement and latent trait parameters fixed at values estimated from male data produced χ^2s approaching those obtained by estimating parameter values from the female data being modeled. The male model was always superior to the female model when adjustments for parsimony were introduced. The inference from these findings is that gender differences may affect that absolute level of performance, but not the relationship between strength and performance.[1] Myers,Gebhardt, Crump, and Fleishman (1993) also used confirmatory factor analysis to demonstrate that males and females produce similar models for strength, so that element of the study represents a replication of a prior finding. The demonstration of comparable associations between strength and performance

is a new finding.

This study also simplified the initial strength-performance model in two ways. First, a lifting performance dimension was replaced by a correlated residual. Second, a weak correlation between an ILM strength dimension and performance was eliminated. Both changes can be interpreted as removing the effects of methods factors that affected specific measures, but were irrelevant to assessing overall strength and performance capabilities. This simplification might be expected to decrease the overall fit of the model to the data, but these changes actually had the effect of slightly improving the overall fit of the model.

The final model produced the simplest possible representation of strength and performance. A single general strength dimension was related to a single general performance dimension. The correlation between the latent traits was $r = .953$, indicating that the ability and performance dimensions were virtually perfect predictors of one another. Note, however, that this correlation describes the relationship between general physical strength and general task performance and estimates the true population correlation if these general dimensions were measured without error. None of the strength and performance measures correlated as highly as this because those relationships are estimated using measures that no doubt reflect task-specific elements of strength and performance and include measurement error. However, the model presented here does identify general strength as a useful generic model of human performance capabilities.

The observation that a single dimension of strength is adequate to account for strength-performance relationships perhaps is surprising in light of prior research. That prior research suggests that strength can be divided into subtypes such as static, dynamic, and explosive strength (e.g., Fleishman, 1964; Myers et al., 1993). Robertson and Trent (1985) designed the set of strength measures to assess static and dynamic strength. The surprise, if there is one, lies in the fact that only a single measure of strength is needed to predict job performance. However, recent work suggests that these different types of strength are highly intercorrelated (e.g., $\underline{r} = .87$ to $\underline{r} = .94$ in Myers et al., 1993). The present findings suggest that these distinctions may have little validity as predictors of task performance. The need for this level of specificity in the ability models is debatable. The factors may be reliably identified when attempting to define the internal structure of strength assessments, but may be of limited importance for models relating strength to other variables.

This model is useful for applications such as those envisioned in current modeling and simulation plans for the Department of Defense (Under Secretary of Defense [Acquisition and Tests], 1995). In that context, the present strength-performance model can greatly simplify some modeling problems. For example, this representation of abilities and performance implies that the performance effects of an environmental stressor such as heat, cold, or sleep loss, can be modeled by determining the effects of that stressor on measures assessing general strength.

The present model of strength and performance also has implications for fitness standards assessment. The present model can be a starting

point for identifying appropriate, broadly predictive job-related fitness indicators. Fitness standards intended to ensure job performance must address a wide range of tasks. Assessments that reflect the general strength dimension of the present model should be ideal for this purpose. Factor analyses of fitness tests suggest that commonly used measures such as sit-ups or push-ups are not suitable for this purpose (e.g., Fleishman, 1964; Hogan, 1991; Myers et al., 1993). Studies relating push-ups and sit-ups to Navy task performance often show little or no relationship (e.g., Marcinik, Hyde, & Taylor, 1995; Robertson & Trent, 1985). In some cases, push-ups and sit-ups may be related to task performance, but other measures still prove more effective in predicting task performance (Beckett & Hodgdon, 1987). For example, one might evaluate a set of potential fitness measures to determine which one(s) load most heavily on the general strength factor defined in this study. Based on the present evidence, the measure(s) with the largest loadings would be the most precise indicators of general strength which is, in turn, a nearly perfect predictor of overall physical performance capacity. If no single test was judged adequate, the present model could be used to select a set of tests that would maximize the accuracy of assessment of general strength.

The strength-performance model even is useful when specific task performance is the focus of attention. To begin with, the model provides the basis for an incremental validity test of the need for task-specific models. Lacking such a model, the decision to construct a task-specific model might be based on null hypothesis testing. In this approach, the basic question would be whether the task to be predicted is related to strength. If so, a set of strength measures specific to that task would be selected to represent the strength-performance relationship. The general performance model developed in this study predicts that all tasks will be related to general strength to some extent. This fact combined with effects of sampling variability can be expected to lead to different predictor profiles for different tasks. As a result, it may appear that each task requires a different predictive model even if the present general model were literally true and sufficient. The current structural model can be used to ensure that no task-specific model is adopted until that model has demonstrated incremental validity relative to the present model. The key to developing task-specific models would be the residual correlations between strength measures and task performance measures after removing the covariation attributable to the relationship between the general strength and performance dimensions. The present analyses found only the Arm Pull-Bolt Torque relationship as a large, replicable residual when this approach was employed. The present data, therefore, would have yielded a task-specific predictive equation for only 1 of 15 performance measures. Generalizing to other potential applications, fewer task-specific models should be needed using this approach compared with the apparent specificity that would result using the null hypothesis approach. The effect should be a substantial reduction in the complexity of task-performance modeling.

The preceding example illustrates how the general model can help develop task-specific models where those are appropriate. The general model can be used to isolate the task-specific components of performance as residual variance. Identifying the correlates of that residual variance is equivalent to identifying additional predictor variables for the specific task. The final model, therefore, might consist of one or more predictors included to represent the general strength dimension and one or

more specific measures chosen because they predicted residual variance. This approach to modeling specific tasks might be applied to critical tasks within an occupational specialty to define job-specific fitness requirements as a refinement to general fitness tests.

The preceding comments illustrate that the general strength-performance model is an example of a bandwidth-fidelity trade-off that has been noted in other areas of human behavior (Anastasi, 1985; Funder, 1991). The trade-off arises from the balance between the stengths and weaknesses of the model. The strength of the present model is the ability to predict a wide range of tasks with just a single general strength dimension. The potential weakness of the model is imprecise prediction of individual tasks. The trade-off between the ability to predict many different tasks with moderate accuracy and the ability to predict specific tasks with high precision defines the "bandwidth-fidelity" principle of prediction (Anastasi, 1985; Funder, 1991). This trade-off is needed to underscore the fact that additional work is needed to clearly define the extent of the trade-off in physical task performance. Procedures such as those employed to assess the utility of psychometric 'g' as a predictor of job performance (Ree, Earles, & Teachout, 1994) can be adapted for this purpose.

The preceding comments must be viewed in the context of potential limitations of the study. First, the sample of ability measures was limited by the absence of aerobic capacity measures. Conceptually, strength and aerobic capacity both contribute to work capacity (Hill, 1993). Prior factor analytic evidence indicates that strength and endurance measures define distinct dimensions of individual differences (Baumgartner & Zuidema, 1972; Fleishman, 1964; Hogan, 1991; Myers et al., 1993). Given that strength accounted for an estimated 91% of the general performance variance in this study, endurance may account for the remaining 9%. This speculation suggests that endurance is less important than strength for Navy tasks, but a factor that accounts for 9% of the variance in performance would be a useful component of models. The hypothesis that aerobic capacity is the source of the unexplained variance in performance should be tested in future work.

The task sample may be a second limitation of the study. While the task sample was designed to meet reasonable criteria and was based on a survey of Navy jobs, all of the tasks were relatively brief (cf., Robertson & Trent, 1985). Such tasks are not expected to depend heavily on endurance elements of fitness. The fact that nine tenths of the general performance variance can be explained by strength differences bears this expectation out. However, this observation may not generalize to all Navy tasks, particularly where a task must be performed repetitively. Beckett and Hodgdon (1987) found that endurance measures were important for predicting performance on a box-carrying task that required performance over a longer period of time than the tasks studied by Robertson and Trent (1985). A wider sampling of the task domain, particularly an extension to tasks of longer duration, therefore would be expected to yield at least a two-dimensional model on both the abilities and task sides of the equation. The result would be a model that paralleled the approach taken by Vogel et al. (1980) to establish fitness requirements for individual jobs. However, it is also fair to note that the procedures used in the construction of Robertson and Trent's (1985) task battery suggest that aerobically demanding tasks may be less common in the U.S. Navy than tasks demanding

strength. Assuming this to be true, the present study provides a basis for modeling the most important component of physical ability relative to U.S. Navy tasks.

The published correlation matrices may be a third limitation of the study. Because of missing data, the correlation coefficients in these matrices were based on different subsets of individuals in the overall samples. This variability in the effective samples for different correlations is a potential source of inconsistencies among the correlation coefficients. This "sampling variability" should be one contributing factor affecting the size of the residuals in the model. The implication is that the GFI values may be lower than would be obtained in a study with complete data on all subjects.

Vickers' (1995) ability-performance model passed the generalizability test in this paper. Model revisions were introduced that further simplified the ability-performance representation. This simple general model provides a working first approximation for representing human physical performance capacities where physical performance is a concern. The model obviously identifies a generic component of individual differences in ability that are relevant to a wide range of tasks and jobs. As such, the model is one step toward the goal of providing sound representations of human abilities for DOD models (Under Secretary of Defense [Testing and Acquisition], 1995). The preceding comments illustrate several ways in which the model can be a starting point for a refined understanding of task performance and physical fitness in Navy jobs. This model can be a point of departure for developing more focused scientific ability-performance models using the analytic tools employed in this study. Properly refined, the model presented can be a job analysis tool for honing our understanding of relationships between physical ability and job performance in the U.S. Navy to provide a stronger basis for establishing and monitoring physical readiness in the Fleet.

Footnotes

[1]The fact that a single ability-performance measurement and predictive model applies to men and women should not be interpreted as a claim that men and women will perform at the same level on a given task. The model accounts for observed covariations between ability and performance. The covariations of interest are computed as moments about the mean (i.e., the mean value is subtracted from each score before the statistic is computed). Thus, mean differences in the level of performance for men and women, which were quite evident in Robertson and Trent's (1985) data, did not figure in the development of the present model. Several types of evidence make it reasonable to expect males and females to yield comparable models when mean differences in performance are disregarded. Men and women generate comparable forces when strength measures are adjusted for various elements of stature, including cross-sectional area, muscle volume, and limb length. Performance differences, therefore, may be a function of differences in body size. Cumulatively, these factors will determine lean body mass, a factor that may be fundamental to differences in strength (Shephard, Bouhlel, Vandewalle, & Monod, 1988). Male-female differences in performance, therefore, may be largely a function of differences in body size.

The validity of the present model does not depend on the accuracy of claims that strength and, by extension, performance are determined by body size. Substantial male and female differences in underlying processes contributing to performance on strength assessment tasks and job performance tasks may exist. Differences in body structure may lead to different strategies or biomechanics in men and women. However, the strategic differences may be equally applicable to the strength assessment tasks and the job performance tasks. The performance measures actually treat people as black boxes and focus solely on output, not how that output is achieved. The present findings indicate that one category of outputs predicts another category of outputs the same way in men and women. This outcome could occur even if the processes within the "black boxes" comprising these two categories were quite distinct. For example, further analysis might demonstrate that the output per unit of lean body mass was different for men and women. If so, one might infer that differences in work strategies, biomechanics, biochemical processes, were present in the two populations. Such male-female differences would not invalidate the present model because those differences occur at a different level of analysis. That level of analysis will be relevant for some purposes (e.g., how to develop programs to enhance fitness), but the current level of analysis is appropriate for the present purpose.

References

Anastasi, A. (1985). Some emerging trends in psychological measurement: A fifty-year perspective. Applied Psychological Measurement, 1, 121-138.

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

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 (Tech. Rep. No. 87-26). San Diego, CA: Naval Health Research Center.

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.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. NY: Academic Press.

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

Funder, D. C. (1991). Global traits: A neo-Allportian approach to personality. Psychological Science, 2, 31-39.

Gaesser, G. A., & Wilson, L. A. (1988). Effects of continuous and interval training on the parameters of the power-endurance time relationship for high-intensity exercise. International Journal of Sports Medicine, 9, 417-421.

Hill, D. W. (1993). The critical power concept: A review. Sports Medicine, 16, 237-254.

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

Jenkins, D. G., & Quigley, B. M. (1992). Endurance training enhances critical power. Medicine and Science in Sports and Exercise, 24, 1283-1289.

Jenkins, D. G., & Quigley, B. M. (1993). The influence of high-intensity exercise training on the W_{lim}-T_{lim} relationship. Medicine and Science in Sports and Exercise, 25, 275-282.

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. Psychological Bulletin, 111, 490-504.

Marcinik, E. J., Hyde, D. E., & Taylor, W. F. (1995). The relationship between the U.S. Navy fleet diver physical screening test and job task performance. Aviation, Space, and Environmental Medicine, 66, 320-324.

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.

Poole, D. C., Ward, S. A., & Whipp, B. J. (1990). The effects of training on the metabolic and respiratory profile of high-intensity cycle ergometer exercise. European Journal of Applied Physiology, 59, 421-429.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. Journal of Applied Psychology, 79, 518-524.

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

Shephard, R. J., Bouhlel, E., Vandewalle, H., & Monod, H. (1988). Muscle mass as a factor limiting physical work. Journal of Applied Physiology, 64, 1472-1479.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. NY: McGraw-Hill.

Stevenson, J., Bryant, T., Greenhorn, D., Deaking, 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.

Under Secretary of Defense (Acquisition and Technology). (1995). Modeling and simulation (M&S) master plan. Washington, DC: Department of Defense.

Vickers, R. R., Jr. (1995). Physical task performance: Complexity of the ability-performance interface (Tech. Rep. No. 95-30). San Diego, CA: Naval Health Research Center.

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. No. 5). Natick, MA: U.S. Army Research Institute of Environmental Medicine.

Wothen, 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. Maximum force exerted was recorded.

Arm Ergometer: Subject turned the wheel on a Monark ergometer as rapidly as possible for 30 s with handle arms set at 4.5 inches and resistance at 600 KPM. Work performed during the exercise period was recorded.

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-lb 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. Maximum weight lifted was recorded.

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. Maximum weight lifted was recorded.

Performance Tasks

Drop-Tank Carry: A gripping device that simulated a tail fin of a drop-tank was attached to a weight of 100 lb. Using the device as a handle, the weight was carried 100 ft in one direction, then 100 ft back to original position after about a 30-s rest. Time for completion was recorded.

Tow-Bar Run, Clear: An aircraft nose gear tow bar with a weight of 62 lb at the grip point was carried or pulled 300 ft. Time to complete the task was recorded.

Tow-Bar Run, Cable: Same tow-bar equipment as immediately above was carried or pulled 300 ft, but the tow bar now had to be taken over 1.5-in pipes simulating aircraft carrier arresting cables. Time to complete the task was recorded.

Fuel Probe Carry: An object with a cylindrical base (12.5 in diameter; 2 in depth) was carried for 50 ft, rest 30 s, then returned to the starting point. A weight of 50, 69, 88, 114, 120 lb was selected by subject as

heaviest with which he/she believed he/she could perform the task. Carrying time and weight were recorded to estimate work rate.

Crucible Pour: Using handles, a simulated crucible was slid 20 ft along a track walking/stepping sideways. The crucible then was returned to initial position stopping every 2 ft to rotate the handles 45 degrees to simulate pouring. Weights for the crucible load were 99, 130, 153, or 168 lb. Each subject chose the maximum weight he/she thought he/she could successfully manipulate in the task. Time and weight were recorded and combined to estimate work rate.

5-Gallon Can Carry: A 5-gal can was carried 170 ft over level surfaces and up and down 2 inclined (not vertical ladders). Load in the can was 0, 35, 45, 60, 75, or 95 lb with the subject choosing the heaviest weight he/she felt he/she could carry. Time and weight were recorded and combined to estimate work rate.

Equipment Carry: Carry a weight with a handle to simulate carrying tool or weapons system component. A weight of 70 lb or 119 lb was chosen by subject and carried 110 ft on level surface, and up and down a ladder. Time for the carry was recorded.

Acetylene Bottle Carry: A gripping device was attached to a cart designed to ride on tracks. Subject held the gripping device then carried the device up 7 steps of a ladder. Loads for the cart could be 88, 106, 133, or 150 lb chosen by the subject as the maximum he/she believed he/she could carry. Time for the carry was recorded.

Mark 82 Bomb Loading: A loaded weight bar was 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 lb. The weight lifted was increased until subject could not lift next highest weight, but could repeat the value just completed.

Canopy Raise, 1-Arm: A canopy-raise simulator was lifted with one hand and a safety strut was inserted. This task was performed while standing in fixed inset steps simulating those found in the side of fighter planes. Weight of the canopy simulator was adjusted from 22, 32, 54, 65, 76, 87, 98 lb to determine the greatest weight the participant could raise.

Canopy Raise, 2-Arm: This task was the same as the 1-arm canopy raise, except that both arms could be used to lift with the safety strut held in one hand during the lift. Maximum weight lifted was recorded.

Rope Pull, Initiating Force: A 25 ft rope was attached to a resistance device which was set at 160 lb. The rope then was pulled 10 ft as rapidly as possible. Time for the pull was recorded.

Rope Pull, Sustaining Force: A 25 ft rope was attached to a resistance device set at 60 lb. The rope then was pulled 20 ft as rapidly as possible. Time for the pull was recorded.

Cart Pull, Initiating Force: Using a handle bar grip attached to same resistance device used in rope pull, the handle was pulled 30 ft with resistance set at 75 lb. Time for the pull was recorded.

Cart Pull, Sustaining Force: Using a handle bar grip attached to same resistance device used in rope pull, the handle was pulled 100 ft with resistance set at 45 lb. Time for the pull was recorded.

Fuel Hose Drag: Using handle bar grip with a resistance device set at 105 lb, participants pulled the handle 80 ft. Time for the drag was recorded.

Power Cable Rig: Using a grip device simulating a 3 ft diameter, 80 lb segment of power cable attached to a resistance device set at 100 lb, participants lifted and pulled the device 40 ft. Time for the pull was recorded.

Bolt Torque: Using a resistance device to assess the torque generated, simulate turning a wrench, participant pulled on the handle of the device with one arm braced against an upright support. The maximum force generated in the pull was recorded.

Appendix B
Parameter Estimates for Final Model for Females and Males Analyzed Separately

The final model reported in Figure 1 of this paper shows parameter estimates from simultaneous analysis of the male and female correlation matrices. The parameters were estimated under the constraint that a single set of parameter values be used to reproduce the correlation matrices for males and females. A comparison of male and female models may be of interest at some time in the future, so the results obtained when the model was estimated separately for males and females are provided here.

The various parameter values differed somewhat across samples. Cumulatively, the differences were large enough to be statistically significant. The χ^2 for the gender-invariant model was substantially greater than the sum of the χ^2s for the gender-specific models (χ^2 = 213.17, 54 df, $p < .001$). Thus, it was reasonable to search for modifications that could improve on the gender-invariant model.

The largest differences between the male and female parameter values were identified to determine the most likely exceptions to the generalization that a single model applied to males and females. The largest difference was the factor loading for the Rope Pull 60 variable on the performance factor (females = .734, males = .338). The second largest difference was the factor loading for the Power Cable Rig variable on the performance factor (females = .531, males = .789). The effects of removing the equality constraints for these two parameters, therefore, were examined.

The effect of removing the constraints on the selected parameters was investigated by retaining a fixed model for all other parameters and freeing the selected parameters sequentially. The first step of freeing the Rope Pull 60 factor loading reduced the model χ^2 by 18.56. Freeing the Power Cable Rig ILM factor loading reduced the χ^2 by a further 9.75. The ATLI for the completely constrained model was .717 (TLI = .738) compared with .719 (TLI = .742) after removing the first constraint. The ATLI was still .719 after removing the second constraint (TLI = .742).

There were two reasons for stopping the search for model modifications at this point. First, the residual χ^2 difference was nonsignificant after freeing the rope pull and power cable rig parameters (χ^2 = 184.86, 52 df, $p < .078$). Second, if it is reasonable to assume that the remaining parameters would have produced smaller improvements in fit, the ATLI would remain constant or decrease with the addition of further freely estimated parameters. The maximum improvement on the ATLI obtained by removing equality constraints, therefore, would be from .717, the value obtained for the fully constrained model, to .719 for the partially constrained model. Clearly, even the largest differences between men and women produced only modest overall changes in the fit of the model.

Table B-1

Final Model Parameter Values for Separate Analyses of Female and Male Correlation Matrices

	Female Data		Male Data	
Ability	***'g_s'***	***ILM***	***'g_s'***	***ILM***
Arm Pull	.711		.775	
Arm Lift	.597		.708	
Arm Ergometer	.670		.831	
ILM Jerk	.637	.603	.624	.760
ILM Press	.666	.723	.706	.644
ILM Elbow	.624	.352	.566	.575
Performance	***'g_p'***		***'g_p'***	
Drop Tank Carry	.640		.522	
Tow-bar Clear	.486		.610	
Tow-bar Cable	.465		.488	
Acetylene Bottle Carry	.699		.472	
5-gallon Can Carry	.664		.579	
Equipment Carry	.697		.576	
Bomb Loading	.575		.729	
Canopy Raise 1	.415		.574	
Canopy Raise 2	.474		.531	
Rope Pull 160	.636		.519	
Rope Pull 60	.744		.338	
Cart Pull 45	.711		.680	
Fire Hose Drag	.643		.679	
Power Cable Rig	.531		.789	
Bolt Torque	.686		.678	
Ability with Performance				
'g_s' - 'g_p' Correlation	.946		.955	
Correlated Residuals				
(1) Canopy Raise 2/ Canopy Raise 1	.443		.388	
Tow-Bar Clear/Tow-Bar Cable	.404		.166	
(2) Arm Pull/Bolt Torque	.205		.268	
(3) Acetylene Bottle Carry/ Drop Tank Carry	.087		.208	
Acetylene Bottle Carry/ Rope Pull 160	-.132		-.277	
Canopy Raise 2/Bomb Load	.104		.115	
Rope Pull 160/Power Cable Rig	.271		.059	
Rope Pull 60/Fire Hose Drag	.131		.130	

Note: (1) = pair of residuals added to reflect variance due to specific task. (2) = residual added to represent ability-performance relationships. (3) = set of residuals added based on post hoc replication across the male and female correlation matrices. See text for details.

The full set of results exploring potential gender differences in the performance model suggests that although the overall χ^2 increase introduced by constraining the factor loadings to be equal for men and women is significant, little is gained by introducing gender-specific parameter estimates. Even relaxing the constraint on the two parameters with the largest absolute differences between men and women produced only slight GFI gains. After making those changes, the residual misfit of the model was within the range expected by chance. However, some researchers may wish to test the possibility that the parameter differences between men and women are reliable across samples even though small. Table B-1 provides the parameter estimates for the final model derived from the male and female correlation matrices. Direct tests of the fit of the gender-specific models to new data sets can be conducted with those parameters.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
	4 Apr 1996	Interim Jan 96 - Mar 96

4. TITLE AND SUBTITLE
Generalizability Test of a Physical Ability-Job Performance Model

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. 96-16

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)*

A strength-performance model developed from male data was generalized to female data reported by Robertson and Trent (1985). Structural equation modeling demonstrated that: (1) male and female correlation matrices could be represented by a single model, and (2) the initial male model could be simplified from two strength dimensions and two performance dimensions to one strength dimension and one performance dimension. Those strength and performance dimensions were highly correlated (r = .953). The specific model developed here may apply only to physical tasks of relatively brief duration, but this first approximation can be useful because it covers a wide range of common Navy physical tasks. Structural modeling provides a tool for refining this initial model to represent a wider range of tasks or to produce higher fidelity in the prediction of specific tasks.

14. SUBJECT TERMS			15. NUMBER OF PAGES
physical tasks	strength	job performance	32
physical fitness	modeling	females	16. PRICE CODE
	structural equation modeling	military	

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