

## ESTIMATING TESTING IMPORTANCE OF TASKS BY DIRECT TASK FACTOR WEIGHTING

Authors: William J. Phalen, Air Force Human Resources Laboratory  
Walter G. Albert, Air Force Human Resources Laboratory  
Darryl K. Hand, Metrica, Inc.  
Martin J. Dittmar, Metrica, Inc.

### INTRODUCTION

This paper is one of a series of presentations delivered at the current and previous two Military Testing Association Conferences to document R&D of an automated, task-data-based outline development procedure for Air Force Specialty Knowledge Tests (SKTs). A companion paper to this one (Albert & Phalen, 1990) provides a brief description of the automated test outline (ATO) procedure. This paper will focus on that part of the ATO procedure having to do with the selection process by which 150 to 250 tasks are selected from a job inventory containing up to 2,000 tasks for inclusion in a Testing Importance Survey booklet. Up to now, rule-based screening procedures have been used to identify potentially important tasks to include in the survey, with cutoffs on percent of members performing each task at the E-5 and E-6/7 paygrade levels and on the recommended training emphasis index being the primary selection criteria. A little over a year ago, research was initiated to derive and validate a minimal subset of regression equations for predicting the SME-furnished testing importance ratings in 28 AFSs with linear combinations of five task-level predictor variables, i.e., percent of members performing (PMP), percent time spent by members performing (PTM), average paygrade of members performing (AG), task learning difficulty (TD), and field-recommended task training emphasis for first-termers (TE). So far, it appears that possibly one, but not more than three, generalized regression equations may adequately classify tasks into their appropriate testing importance categories. These equations will, hopefully, perform several important functions. First of all, they should provide a more accurate and defensible task selection procedure for surveying AFSs that have not been previously surveyed. Secondly, the predicted testing importance (PTI) values generated by the equations should be able to serve as surrogate testing importance indices when time or budget constraints prevent the administration of testing importance surveys. Thirdly, when a new job inventory is developed and administered in an AFS whose testing importance data are based on the old job inventory tasks, the new data for the predictor variables should be available to use in conjunction with one of the generalized regression equations to generate PTI values for all the tasks in the new job inventory.

But the application of these PTI equations also raises several pertinent questions: (1) How can we determine which PTI equation should be used to generate PTI values for a previously unsurveyed AFS? (2) Can SMEs provide direct estimates of AFS-specific weights for the five predictor variables that are nearly as accurate for an AFS as the generalized regression weights? (3) Is it possible that the need for regression-generated or SME-derived weighting is obviated by simple unit weighting of the five predictor variables? The potential value of direct estimation of predictor weights by SMEs was anticipated back in 1987; accordingly, an SKT Task Factor Testing Importance Survey booklet was developed and administered to the SMEs in all AFSs for which SKTs were developed in 1988, 1989, and 1990 (to date). The booklet used in 1988 contained seven factors, the two additional ones being "consequences of inadequate performance" (CIP) and "requirement for prompt performance" (RPP), the latter being a rewording of the old "task delay tolerance" factor in order to reverse the direction of the scale and make it consistent for all factors. In 1989, it was decided to limit the task factors surveyed to the five which were routinely surveyed by the USAF Occupational Measurement Squadron (USAFOMS); thus, CIP and RPP were dropped. The elimination of the CIP and RPP factors also made it possible to assess the effect of their presence or absence on the other five factors. In 1990, the CIP and RPP factors were restored to the survey in order to introduce more variance into the profiles of the SME-furnished factor weights and thus eliminate some fuzziness from the clustering solution. The availability of data on the same seven factors for the same AFSs in 1988 and 1990 made it possible to assess the stability of factor weights over a two-year period, assuming, of course, that the SMEs in both periods were equally representative of their AFSs.

## THE SURVEY INSTRUMENT

The SKT Task Factor Testing Importance Survey is administered to all SMEs who have been sent by their respective commands to participate in the development of SKTs in their AFSs. To date, approximately 1,000 SMEs have been surveyed. The survey is group-administered by a member of the USAFOMS test development staff immediately following the SKT in-briefing. It takes about 10 minutes to read the instructions, fill in the background section, and provide ratings on the seven listed factors (1 to 7 scale). In order to clearly communicate what the SKT task factor rating process is all about, the rating instructions, scale, and factor definitions as they appear in the survey booklet are shown in Figure 1.

## RESULTS

- A. Reliability Analysis. There were 35 AFSs in which the SKT Task Factor Testing Importance Survey was administered in 1988 and again in 1990. In most instances, no SMEs appeared in both survey samples. As shown in Table 1, the average number of raters per AFS in 1988 was 3.50, and in 1990, the average was 3.59. The average correlation between the mean factor profiles (across seven factors) for the 35 AFSs was .4841 (correlations averaged through  $Z_r$ ). A value this high was considered very acceptable, especially since it involved a two-year time interval between administrations and small numbers of different raters per AFS at both points in time. This value compares very well with the average test-retest reliability of .5835 that was obtained on task-level testing importance ratings for 26 raters in 20 AFSs with a 3-to-4-month interval (Weissmuller, Dittmar, & Phalen, 1988). These raters were surveyed by mail and were later surveyed again when they were selected to serve on an SKT development team. The difference between the two reliability coefficients was found to be nonsignificant ( $p = .4337$ ). As a further test, the 1988-to-1990 factor profile correlation ( $\bar{r} = .4841$ ) was treated as a group measure of interrater reliability ( $R_{kk}$ ) with no time interval involved, and the  $R_{kk}$  was reduced to a single-rater reliability value ( $R_{11}$ ) for comparison with the mean  $R_{11}$  value for task-level testing importance ratings across all 28 AFSs that had been surveyed. The computed  $R_{11}$  value for a composite reliability ( $R_{kk}$ ) of .4841 based on an overall average of 3.54 raters per factor profile was .2649. The average  $R_{11}$  for the task-level testing importance ratings across the 28 surveyed AFSs was .2640, an almost identical value. Yet, the former involved a two-year interval and the latter is a concurrent measure of internal consistency.
- B. Relative Weighting of Common Task Factors in the Five- and Seven- Factor Sets. Two tests were applied to determine whether the relative weights of the common five factors were affected by adding or removing the additional two factors (i.e., CIP and RPP). In the first test, each factor was given an overall rank in terms of its mean rating in 1989 (five-factor survey) and its mean rating in 1988 and 1990 separately (seven-factor surveys). The Mann-Whitney test was applied to assess the differences in the sums of ranks. The mean ratings of the PTS, AG, and TD factors were relatively unaffected by the presence or absence of the additional factors, but PMP and TE showed significant shifts in their mean ratings ( $p < .01$ ). Both were significantly higher when CIP and RPP were absent (or significantly lower when CIP and RPP were present). A test was also applied to determine whether the sizes of the differences between the PMP and TE means in the five-factor vs. the seven-factor environment were related to the sizes of the mean CIP and RPP values. Regression equations of the form  $\overline{PMP}_7 - \overline{PMP}_5 = W_1 \overline{CIP}_7 + W_2 \overline{RPP}_7$  were applied. None of the regression results were found to be significant. Thus, while it can be said that  $\overline{PMP}$  and  $\overline{TE}$  were affected in a given direction by the presence or absence of  $\overline{CIP}$  and  $\overline{RPP}$ , there was no indication that the level of difference was proportional to the level of  $\overline{CIP}$  and  $\overline{RPP}$ .

## SECTION II. INSTRUCTIONS

Imagine that you have been asked to review the job-task statements in the most recent USAF Job Inventory administered in the career field for which you are developing SKTs. This survey could contain anywhere from 500 to 1200 or more task statements. Next, assume that you have been asked to rate each task statement indicating how important it is to include the job knowledges needed to perform that task on a Specialty Knowledge Test. A task would be rated high in testing importance if it requires knowledges that are critical to successful job performance within the career field.

You are in luck, however. You are not being asked to provide these 500 or more ratings. Instead, seven factors (or types of information) have been proposed as possible factors in determining the testing importance level of a task. These seven factors, along with their descriptions, are shown in Section II, SKT TASK FACTOR TESTING IMPORTANCE RATING SCALE. You are asked to rate each task factor on how important it is to consider this factor when assigning a testing importance rating to the tasks performed by airmen in the Air Force Specialty for which you are developing SKTs. Using the scale provided, determine the most appropriate rating and record your rating in the column provided.

### SECTION II: SKT TASK FACTOR TESTING IMPORTANCE RATINGS

#### RATING SCALE FOR FACTORS IN TESTING IMPORTANCE

This factor has:

- 7 = Extremely High Importance
- 6 = High Importance
- 5 = Above Average Importance
- 4 = Average Importance
- 3 = Below Average Importance
- 2 = Low Importance
- 1 = No Importance

<u>Rating</u>	<u>Factor</u>
_____	1. <u>Percent Members Performing</u> : a measure of the proportion of all airmen who perform the task.
_____	2. <u>Average Percent Time Spent</u> : a measure of the proportion of the total work time that airmen in the AFS spend performing the task.
_____	3. <u>Average Grade</u> : the average grade of all airmen who perform the task.
_____	4. <u>Learning Difficulty</u> : a measure of the relative length of time required to learn to perform the task properly.
_____	5. <u>Consequences of Inadequate Performance</u> : a measure of the probable seriousness of failing to perform the task properly. The impact is measured in terms of possible injury or death, damage to equipment, wasted supplies or lost work-hours, etc.
_____	6. <u>Requirement for Prompt Performance</u> : a measure of the length of time from the moment that an airman is aware that a task will need to be done up to the point at which the task MUST be performed. In other words, does the airman have to be able to perform the task immediately, or does he or she have time to consult a manual or seek guidance?
_____	7. <u>Field-Recommended Entry-Level Training Emphasis</u> : a measure of how strongly NCOs in the field have recommended the task for inclusion in formal, structured training programs for entry-level airmen. Structured training may include resident technical school, on-the-job training (OJT), field training detachments (FTDs), or career development courses (CDCs).

Figure 1. SKT Rating Form

C. Clustering of Factor Profiles vs. Clustering of PTI Regression Equations. One objective of gathering task factor ratings from SMEs was to provide a means of determining which one of several generalized regression equations should be applied to previously unsurveyed AFSs to select the appropriate set of tasks for inclusion in a Task Testing Importance Survey. If AFS factor profiles produced a clustering of AFSs that corresponded to the clustering of AFSs on similarity of regression equations, then regression equation group membership could be defined for task factor clusters of AFSs for which there were no regression equations. Various attempts were made to produce corresponding clustering solutions, but no adequate match could be generated. A major impediment was the fact that even in the case in which the input sample of factor profiles contained the maximum amount of variance (1988, 1989, and 1990 combined) the "between" overlap for the last two groups to merge was 86.3% and the total sample "within" overlap was 93.2%. On the other hand, the clustering of regression equations did not seem to indicate a need for more than one equation. Thus, a lack of variance was present in these data, as well. If additional research indicates that only one overall regression equation is needed for all AFSs, then the need for a procedure to select the appropriate regression equation for a previously unsurveyed AFS vanishes.

D. Comparison of Regression- vs. Factor-Weighted Equations for Predicting Testing Importance of Tasks. Table 1 shows the predictive efficiency of the AFS-specific PTI regression equations for 25 AFSs for which task-level testing importance indices were available and for which SMEs had provided factor weights in 1988, 1989, or 1990. Since the derivation and validation of the regression equations and their predictive efficiency are discussed in detail in a companion paper (Albert & Phalen, 1990), the correlations of predicted and actual testing importance values for the 25 AFSs are reported here only for their comparison with the correlations produced by the SME-based factor-weighting approach (which standardizes each task factor before applying the factor weights and sums the cross-products into a testing importance composite). In Table 1, only the highest correlations computed for the 1988, 1989, and 1990 factor weights and all possible combinations thereof are reported in order to show the highest correlations this approach can hope to produce for comparison against the best alternative, i.e., the least-squares fit of task-level indices for the five task factors (predictors) to the indices of task-level testing importance (criterion).

For some unexplainable reason, the 1990 factor weights uniformly produced lower correlations than the 1988 weights. Overall, the factor-derived correlations averaged to a respectable  $\bar{r} = .602$  at the E-5 level and  $\bar{r} = .606$  at the E-6/7 level, compared to  $\bar{r} = .798$  and  $.786$  for the E-5 and E-6/7 regression-derived correlations, respectively. The difference is significant ( $p < .01$ ) in both cases, but the real difference is in the lack of uniformity of fit of the factor-derived approach; i.e., in some cases, it matches the regression-derived correlations quite well, and in other cases rather poorly. It appears that the SME-furnished factor-weighting approach is not an acceptable alternative to the regression approach, as long as the regression alternative remains supportable.

E. Differential vs. Unit Weighting of Factors. Because there was little variance in the SME-derived factor weights, and substantial positive correlations existed between the five task factors and the testing importance criterion, with the exception of average grade (Weissmuller, Dittmar, & Phalen, 1989), there was a distinct possibility that a unit-weighted linear composite of the standardized task factors might do almost as well as the differentially weighted composite. The effect of unit weighting on the correlations with testing importance are shown in Table 1 under the heading "Unit." The unit weighting approach produced correlations for both the E-5 and E-6/7 levels that were generally close to the correlations derived from differential weighting by SMEs, with only two instances showing a substantial drop in correlation (both within the same AFS); but 14 correlations based on unit weighting were actually higher than those based on differential weighting. Tests of significance of difference between the  $\bar{r}$ 's for differential and

TABLE 1. NUMBER AND PREDICTED TESTING IMPORTANCE CORRELATIONS FOR 25 AIR FORCE SPECIALTIES

Air Force Specialty (AFS)		Factor Importance Number of Raters			Number of Tasks in TI Survey	E-5 r = TI with PTI			E-6/7 r = TI with PTI		
		1988	1989	1990		Regr.	SME**	Unit	Regr.	SME**	Unit
112X0	Inflight Refueling	4	2	5	127	.909	.734	.698	.903	.772	.740
121X0	Survival Training	3	4	2	148	.908	.717	.497	.824	.725	.564
231X0	Visual Information Media	4	5	4	109	.640	.528	.488	.645	.469	.440
241X0	Safety	—	3	5	162	.827	.766	.763	.807	.727	.722
274X0	Command and Control	3	3	4	147	.830	.555	.496	.798	.631	.594
275X0	Tactical Air Command and Control	2	2	—	198	.862	.757	.703	.865	.816	.768
309X0	Space Systems Maintenance	—	4	3	214	.626	.262	.255	.603	.347	.349
392X0	Maintenance Scheduling	4	4	—	149	.795	.873	.865	.791	.635	.627
454X3	Aircraft Fuels	—	4	2	155	.794	.716	.636	.709	.707	.667
454X4	Aircraft Pneumatic Systems	3	4	2	179	.825	.564	.559	.828	.600	.604
456X2A	Defensive Fire Control Systems	2	3	2	165	.863	.759	.754	.884	.639	.601
456X2B	Defensive Fire Control Systems	6	2	3	145	.769	.667	.693	.732	.534	.537
457X1	Helicopter Maintenance	2	3	—	187	.789	.227	.268	.834	.402	.439
458X3	Fabrication and Parachute	3	1	4	175	.845	.264	.229	.778	.288	.270
472XX	Special Vehicles Maintenance	4	5	6	257	.830	.489	.390	.849	.588	.520
542X1	Electrical Power Lines	—	—	4	198	.862	.554	.505	.848	.680	.655
552X0	Structural	5	2	4	202	.757	.362	.327	.684	.498	.471
566X0	Pest Management	4	2	4	168	.698	.569	.509	.720	.642	.631
674X0	Cost Analysis	3	—	—	148	.773	.497	.505	.770	.582	.588
753X0	Combat Arms	—	4	2	222	.669	.659	.662	.712	.701	.700
791X0	Public Affairs	—	3	—	147	.720	.670	.620	.716	.359	.465
791X1	Radio and Television Broadcasting	4	—	—	156	.713	.497	.432	.668	.580	.589
908X0	Environmental Medicine	4	3	—	192	.512	.299	.320	.480	.356	.370
914X0	Mental Health	—	3	5	140	.908	.647	.537	.858	.674	.580
915X0	Medical Materiel	5	4	—	190	.786	.718	.735	.874	.740	.735
		3.50*	3.18*	3.59*	171.20*	.798*	.602*	.565*	.786*	.606*	.582*

\* Mean for column

\*\* r = MAX (1988, 1989, 1990)

unit weighting at the E-5 and E-6/7 levels (.602 vs. .565, and .606 vs. .582, respectively) yielded no significant differences. These findings clearly indicate that there is virtually nothing to be gained by continuing to gather factor importance ratings from SMEs, since unit weighting of the factors is equally effective.

## DISCUSSION

The findings of this study suggest one positive conclusion and three negative conclusions. The positive conclusion is: (1) Factor importance weights display good reliability, even when the interval between administrations is as long as two years. The negative conclusions are: (1) The factor importance weighting approach does not yield correlations with task-level testing importance that would permit abandonment of the more rigorous regression approach, which requires the administration of task-level testing importance surveys in order to obtain criterion data for generating a least-squares solution. (2) There does not appear to be sufficient variance in the profiles of factor weights to provide a clustering of AFSs that corresponds sufficiently well with the clustering of AFS-specific regression equations; therefore, the clustering of profiles of factor weights is not useful for indicating which generalized regression equation should be used for a particular AFS (assuming that more than one equation will be needed to adequately cover all AFSs). (3) Since unit weighting of the testing importance factors is virtually as good as SME-furnished differential weights, there is little to be gained by continuing to gather factor importance ratings from SMEs.

## RECOMMENDATIONS

Discontinue administration of the Testing Importance Factors Survey and concentrate instead on improving the predictive efficiency and classification accuracy of the regression-based procedure.

## REFERENCES

- Albert, W.G., & Phalen, W.J. (1990). Development of equations for predicting testing importance of tasks. Proceedings of the 32nd Annual Conference of the Military Testing Association, Orange Beach, AL.
- Weissmuller, J.J., Dittmar, M.J. & Phalen, W.J. (1989). Automated test outline development: research findings (AFHRL-TP-88-70, AD-215 401). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.