AFHRL-TP-88-70

AIR FORCE HUMAN RESOURCES LABORATORY

AUTOMATED TEST OUTLINE DEVELOPMENT: RESEARCH FINDINGS

Johnny J. Weissmuller
The Texas MAXIMA Corporation
8301 Broadway, Suite 212
San Antonio, Texas 78209

Martin J. Dittmar
Metrica, Incorporated
8301 Broadway, Suite 215
San Antonio, Texas 78209

William J. Phalen
MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

November 1989

Interim Technical Paper for Period October 1987- February 1989

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

DANIEL L. LEIGHTON, Colonel, USAF
Chief, Manpower and Personnel Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 1989	Interim - Oct 87 to Feb 89

4. TITLE AND SUBTITLE

Automated Test Outline Development: Research Findings

5. FUNDING NUMBERS

PE: 62205F
PR: 7719
TA: 20
WU: 14

6. AUTHOR(S)

Weissmuller, J.J.; Dittmar, M.J.; Phalen, W.J.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

8. PERFORMING ORGANIZATION REPORT NUMBER

AFHRL-TP-88-70

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Paper presented at the 30th Annual Conference of the Military Testing Association, 27 November - 2 December 1988.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Automated Test Outline (ATO) research and development effort was designed to explore and resolve both technical and logistical problems associated with using occupational survey data to derive weighted subject matter areas for Specialty Knowledge Test (SKT) outlines. This paper discusses some of the issues that were addressed and resolved by research accomplished between October 1987 and December 1988. Issues discussed include: (a) the process for selecting appropriate subsets of tasks from a full task inventory for mailout to subject-matter experts in order to obtain task-level testing importance ratings; (b) interrater and test-retest reliability indices for testing importance ratings in 28 Air Force specialties; (c) the validity of the ATO procedure, as measured by the SKT teams' adherence to the computed testing importance weights for each duty-level outline area and each task; and (d) the relationship between field-validated testing importance and a variety of routine available task factors, such as field-recommended training emphasis, task learning difficulty, average grade of members performing, and percent members performing and average percent time spent by members performing at the E-5 and E-6/7 paygrade levels.

14. SUBJECT TERMS

automated test outline
job analysis
occupational analysis
task analysis
test development
test outline

15. NUMBER OF PAGES

14

16. PRICE CODE

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

AFHRL Technical Paper 88-70 November 1989

AUTOMATED TEST OUTLINE DEVELOPMENT: RESEARCH FINDINGS

Johnny J. Weissmuller

The Texas MAXIMA Corporation
8301 Broadway, Suite 212
San Antonio, Texas 78209

Martin J. Dittmar

Metrica, Incorporated
8301 Broadway, Suite 215
San Antonio, Texas 78209

William J. Phalen

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

Lawrence O. Short, Lt Col, USAF
Chief, MPT Technology Branch

Paper presented at the 30th Annual Conference of the Military Testing Association, 27 November - 2 December 1988.

SUMMARY

The Automated Test Outline (ATO) research and development effort was designed to explore and resolve both technical and logistical problems associated with using occupational survey data to derive weighted subject matter areas for Specialty Knowledge Test (SKT) outlines. This paper discusses some of the issues that were addressed and resolved by research accomplished between October 1987 and December 1988. Issues discussed include: (a) the process for selecting appropriate subsets of tasks from a full task inventory for mailout to subject-matter experts in order to obtain task-level testing importance ratings; (b) interrater and test-retest reliability indices for testing importance ratings in 28 Air Force specialties; (c) the validity of the ATO procedure, as measured by the SKT teams' adherence to the computed testing importance weights for each duty-level outline area and each task; and (d) the relationship between field-validated testing importance and a variety of routinely available task factors, such as field-recommended training emphasis, task learning difficulty, average grade of members performing, and percent members performing and average percent time spent by members performing at the E-5 and E-6/7 paygrade levels.

PREFACE

This work was completed under Work Unit 77192014, Research in Manpower and Personnel Technologies, Advanced and Exploratory Development of Occupational Measurement Technology. This paper was presented at the 30th Annual Conference of the Military Testing Association and published in the proceedings of that event.

TABLE OF CONTENTS

LIST OF TABLES

AUTOMATED TEST OUTLINE DEVELOPMENT: RESEARCH FINDINGS

I. INTRODUCTION

This paper reports research findings related to the production and use of automated test outlines (ATOs) for Air Force Specialty Knowledge Test (SKT) construction. Following a short review of the process used to develop ATOs, the major focus will be on the reliability and validity of obtained results.

II. DEVELOPMENT PROCESS

Briefly, task-level Predicted Testing Importance (PTI) values derived from off-the-shelf task factor components of Field Recommended Training Emphasis (TE); Task Learning Difficulty (TD); and Percent Members Performing (PMP), Percent Time Spent (PTS), and Average Grade Performing (AG) at the E-5 and E-6/7 paygrade levels are used to delimit Air Force specialty (AFS) knowledge domains in terms of restricted subsets of tasks from full task inventories. These subsets of tasks are then administered by mailout approximately 3 to 4 months prior to the start of SKT construction projects to random samples of 50 to 70 senior noncommissioned officers (NCOs) who currently work in the AFSs. These NCOs rate each task within the subset on a 7-point scale of specialty knowledge testing importance. The resulting field inputs are then processed, analyzed, and subsequently used to determine testing importance (TI) weights for each task in the mailout and to calculate test outline weights (numbers of test items to be written) for each major duty area of the specialty. These task-level testing importance ratings by field NCOs are key components within the ATO development process.

III. RELIABILITY AND VALIDITY ESTIMATES

SKTs are normally developed 6 to 18 months before their scheduled administration to Air Force enlisted personnel. Consequently, direct reliability and validity estimates for those SKTs constructed from ATOs will not be available until after the test administration cycle is completed (early 1989). However, the "goodness" of the process used to generate ATOs can be evaluated by examining reliability and validity indices associated with a primary component of the process: Field Validated Testing Importance (FVTI).

To date, ATOs have been developed for 28 AFSs. Table 1 lists interrater reliability estimates of FVTI for each of these specialties.

The Table 1 data show reasonably good interrater reliability estimates (all $R_{11} = 0$ probabilities were less than .01).[1] Over 75% of the R_{11}'s exceeded .20 and no R_{kk} estimate was below .75. Although these levels of reliability are in accordance with expectations (as judged by the Air Force's ongoing experiences with the reliability of field NCO ratings of task learning difficulty), future research efforts will be directed toward increasing interrater reliabilities, primarily through clearer and more simplified rating booklet (task subset) instructions and the identification and segregation of more relevant rater subgroups. The low number of raters in some AFSs was attributable to a variety of causes: small rater populations; testing importance survey booklets

[1] Deviant raters (i.e., those whose correlations with the composite of all raters were not significantly better than zero at $\alpha = .05$) were eliminated from the calculation of R_{11} and R_{kk} (usually less than 10% of sample).

returned too late for inclusion in ATO projects; high percentage of senior NCOs on travel, leave, or permanent change of station; and higher-than-average percentage of deviant raters.

Table 1. **FVTI Interrater Reliability**

Specialty		N-Tasks	N-Raters		R_{11}		R_{kk}	
		E-5/6/7	E-5	E-6/7	E-5	E-6/7	E-5	E-6\7
In-Flight Refueling	112X0	127	43	43	.478	.473	.975	.975
Survival Training	121X0	148	31	30	.225	.199	.900	.882
Still Photography	231X2	145	25	26	.168	.148	.835	818
Audiovisual Production	232X0	135	24	24	.295	.323	.909	.920
Safety	241X0	162	45	45	.270	.282	.943	.946
Command and Control	274X0	147	30	30	.294	.282	.926	.922
Aircraft Control and Warning	303X0	152	46	51	.117	.121	.859	.875
Space Systems Maintenance	309X0	214	24	24	.123	.156	.757	.803
Defensive Fire Control	231X1E	145	6	6	.405	.397	.792	.787
Defensive Fire Control	321X1G	165	15	15	.384	.397	.903	.908
Precision Measuring Equipment	324X0	161	29	29	.233	.255	.897	.908
Maintenance Scheduling	392X0	149	18	18	.376	.366	.916	.912
Aircraft Fuel	423X3	155	24	25	.205	.208	.861	.868
Aircraft Pneudraulic	423X4	179	27	28	.164	.223	.840	.889
Fabrication and Parachute	427X3	175	40	41	.281	.229	.940	.924
Helicopter Mechanic	431X0	187	39	41	.228	.228	.938	.942
Special Vehicle Maintenance	472X0	257	30	30	.218	.249	.892	.907
Electrical Power Lines	542X1	198	25	25	.275	.298	.903	.914
Electrical Power Production	542X2	155	40	41	.188	.232	.902	.925
Structural	552X0	202	33	34	.132	.138	.834	.845
Cost Analysis	674X0	148	33	33	.192	.217	.887	.901
Social Actions	734X0B	137	25	24	.245	.281	.887	.901
Combat Arms	753X0	222	11	11	.316	.367	.836	.864
Public Affairs	791X0	147	31	31	.207	.319	.890	.936
Radio and TV	791X1	156	22	22	.293	.273	.901	.892
Environmental Medicine	908X0	192	34	35	.247	.308	.917	.939
Mental Health	914X0	140	25	24	.302	.263	.914	.894
Medical Materiel	915X0	190	31	34	.192	.351	.880	.948

Because the subject-matter experts (SMEs) who develop SKTs are from the same population as those selected to complete testing importance field rating booklets, we were able, in a limited number of cases, to readminister the rating booklets to assess the stability of FVTI ratings over time. Results of this test-retest effort are contained in Table 2.

In approximately 75% of the cases, test-retest correlations were .5 or higher. Although there was a 3- to 4-month period between first (X) and second (Y) administrations and totally different administration environments (X was self-administered at the rater's home station, and Y was administered at the USAF Occupational Measurement Center (USAFOMC) by the contractor as part of the initial SKT construction in-briefing), these correlations in most instances tend to support rating stability. They are 1.5 to 3 times higher than the interrater agreement coefficients (R_{11}'s) in Table 1, except for Air Force Specialty Codes (AFSCs) 321X1E and 427X3. There was no discernible trend with respect to mean ratings between first and second administrations. Rater variance tended to be smaller for the second administration ($p < .02$, Sign-Rank Test of Differences), indicating a conservative rating policy, which may be the result of the more structured second administration environment.

Although the validity of SKTs constructed from ATOs cannot at this time be assessed directly in terms of increased job relevance of test content, it can to some extent be inferred by examining characteristics of FVTI ratings. As previously stated, FVTI ratings are used to establish

Table 2. **Test-Retest Reliability (FVTI)**

AFSC	N-Raters	First administration		Second administration		Avg r_{xy}
		$\bar{X}$	SD	$\bar{Y}$	SD	
12150	1	5.3	1.53	4.3	1.64	.61
12170	1	4.3	1.46	4.0	1.51	.39
32151E	1	4.4	1.29	4.3	.93	.49
32151G	2	4.7	1.40	4.0	1.28	.77
32171E	1	4.4	1.29	4.3	.93	.48
32171G	2	4.7	1.34	4.0	1.29	.76
39250	2	4.4	1.86	4.1	1.95	.77
39270	2	3.9	1.57	3.8	1.85	.66
42353	1	4.4	.61	4.5	.89	.54
42373	1	4.3	.56	4.5	.89	.58
42753	1	4.7	.86	5.2	.64	.43
42773	1	3.9	1.64	4.2	.64	.27
54252	1	3.8	.99	4.2	.64	.45
54272	1	3.9	1.09	4.3	.68	.43
55250	1	5.2	1.13	4.9	.81	.58
55270	1	5.4	1.17	5.1	.86	.58
67450	2	4.3	1.23	4.2	.77	.67
67470	2	4.3	1.27	4.6	95	.61
90850	1	2.5	1.61	2.6	1.43	.79
90870	1	2.5	1.61	2.6	1.43	.79

ATO weights (the recommended numbers of test items to write for each major duty area). For any given AFS, the number of these major duty areas can vary from as few as 8 to as many as 26. It seems reasonable to assume that the extent to which SKT construction teams adhere to these recommended duty area weights is an indication of the SME-judged goodness (validity) of the FVTI ratings and, to a lesser extent, the SME-judged appropriateness of the ATO development process. Table 3 shows the correlations between recommended and final (as adjusted by the test construction team) major duty area weights. An alternative explanation for the high degree of adherence to the ATO weights by the SMEs, in the opinion of several test psychologists, was the flexibility the SMEs had in selecting knowledge requirements when writing an item on a task.

The correlations listed in Table 3 range from .83 (AFSC 32151E) to 1.00. For the total set of AFSs, approximately 88% of the automated outlines had correlations (between recommended and actual weights) of .95 or higher. In approximately 44% of the cases, no weight change was necessary. These are positive indications and speak to the validity of the FVTI ratings.

In addition to being used to calculate recommended major duty area weights, FVTI ratings are also used to differentiate outline tasks into A (high testing importance) through D (low testing importance) categories, depending on the mean FVTI value computed for each task. If FVTI ratings are valid, one would expect a greater percentage of A tasks to be used to generate test items than B, C, or D tasks; to a lesser extent, one would expect that B tasks would be used to generate test items at a somewhat higher rate than C or D tasks, and that D tasks would have the lowest usage rate of all. Table 4 lists by task category (A, B, and C) the ratio of the proportion of test items written on tasks in that category to the proportion of tasks in that category appearing in the E-5 and E-6/7 outlines combined. The D category is not listed, because only one D task was used by one AFS. Even though tasks in the D category were not to be used without written justification, it is nevertheless significant that only one team felt

the need to justify the use of only one D task. On the other hand, although SMEs were required to write a minimum of one item on each A task, it is significant that the item/task ratio for A tasks substantially exceeds 1.0 in all AFSs, and there are no reversals in the expected decrease in item/task ratios from A to B to C.

Table 3. **Correlations Between Recommended and Actual Test Outline Weights Used**

AFSC	Number of outline areas	r	AFSC	Number of outline areas	r
11250	8	1.00	54271	16	1.00
11270	8	1.00	54272	23	1.00
23152	10	.99	55250	23	1.00
23172	10	.98	55270	23	1.00
32151E	12	.83	67450	12	.98
32171E	12	.84	67470	12	.95
39250	17	.96	73450B	12	.96
39270	17	.95	73470B	12	.96
42353	16	1.00	79150	16	1.00
42354	14	1.00	79151	16	.99
42373	16	1.00	79170	16	1.00
42374	14	1.00	79171	16	.99
42753	26	.99	90850	16	.85
42773	26	.99	90870	16	.94
54251	16	1.00	91550	15	.98
54252	23	1.00	91570	15	.99

Table 4. **Task Use Ratio**

AFSC	Task type	N (Tasks)	Item/task ratio	AFSC	Task type	N (Tasks)	Item/task ratio
112X0	A	38	1.9	542X1	A	70	1.9
	B	106	.9		B	134	.8
	C	106	.3		C	129	.7
231X2	A	45	1.9	542X2	A	53	1.1
	B	111	1.2		B	97	.7
	C	112	.4		C	97	.6
321X1E	A	64	2.0	552X0	A	53	2.1
	B	94	.4		B	160	1.1
	C	94	.1		C	166	.4
392X0	A	48	1.8	674X0	A	60	1.3
	B	42	1.4		B	97	.5
	C	156	.6		C	96	.4
423X4	A	11	1.9	908X0	A	47	2.4
	B	83	1.2		B	40	1.7
	C	209	.8		C	161	.5
427X3	A	54	2.6	915XO	A	11	1.3
	B	133	.8		B	40	.7
	C	139	.5		C	189	.7

Tables 5 and 6 examine the relationships between FVTI and five task-level factors: Predicted Testing Importance (PTI), Percent Time Spent by Members Performing (PTM), Training Emphasis (TE), Task Learning Difficulty (TD), and Average Grade Performing (AG). Because of the large number of tasks used to compute these correlations, a coefficient of $\pm .17$ is significant at $\alpha = .05$ (two-tailed), and $\pm .22$ at $\alpha = .01$ (two-tailed).

Table 5. **Correlation of FVTI (E-5) with Other Task Factors**

AFSC	PTI (E-5)	PTM (E-5)	TE	TD	AG	N (Tasks) E-5/6/7
112X0	.55	.28	.72	.49	.00	127
121X0	.69	.13	.81	.34	-.37	148
231X2	.43	.47	.53	.09	-.11	145
309X0	.22	.00	.12	.57	-.03	214
321X1E	.51	.40	.49	.58	-.15	145
321X1G	.80	.44	.73	.65	-.25	165
324X0	.55	-.04	.41	.55	-.01	161
427X3	.72	-.16	61	57	-.38	175
542X1	.83	-.21	.65	.72	.07	198
542X2	.41	-.18	.18	.66	-.02	155
552X0	.53	.07	.71	-.08	-.02	202
674X0	.64	-.08	64	40	.25	148

Table 6. **Correlation of FVTI (E-6/7) with Other Task Factors**

AFSC	PTI (E-6/7)	PTM (E-6/7)	TE	TD	AG	FVTI (E-5 vs E-6/7)
112X0	.37	.38	.58	.61	.21	.97
121X0	.54	.21	.65	.43	-.23	.93
231X2	.33	.27	.29	.34	.14	.89
309X0	.18	.08	-.04	.50	.25	.92
321X1E	.57	.13	.46	.58	-.11	.99
321X1G	.80	.25	.67	.71	-.12	.98
324X0	.54	-.16	.11	.74	.30	.89
427X3	.69	-.11	.40	.65	-.10	.94
542X1	.81	.29	.48	.73	.32	.94
542X2	.45	.05	-.02	.74	.30	.91
552X0	.49	.19	55	.12	.11	.92
674X0	.68	-.07	.55	.43	.35	.95

As can be seen in Tables 5 and 6, FVTI correlations with PTI at both the E-5 and E-6/7 levels are relatively high, the single exception being AFSC 309X0. This AFS is probably the most diverse (heterogeneous) of all those for which outlines were developed. This diversity may also account to some extent for the relatively weak relationship between FVTI and PTI for this AFS. The FVTI correlations with TE and TD are in the expected direction and at the appropriate level for most of the sampled AFSs. At the E-5 level, we would expect TE to have a stronger impact on FVTI than at the E-6/7 level, as TE is essentially a measure of recommended training emphasis for first-term airmen. Conversely, we would expect TD to have a stronger relationship with FVTI at the E-6/7 level than at E-5 level, in that TD is an estimate of how difficult it is to learn to perform a task. Both of these expectations are confirmed by the correlations in Tables 5 and 6, which lend a degree of convergent validity to the testing

importance measure. However, the consistently strong, positive relationship between FVTI (E-5) and FVTI (E-6/7) could indicate the presence of an unwanted autocorrelation resulting from the dual-column "E-5/E-6/7" format employed in the rating booklets used to collect FVTI information.

It is evident from these findings that a single PTI equation for predicting FVTI will not be a feasible objective, and that more attention must be given to TD, which has been underweighted in the procedure for selecting tasks to be rated on testing importance.

IV. CONCLUSION

Although the statistical information gathered thus far is by no means overwhelming, it is very encouraging that it is uniformly in the right direction for almost all AFSs in which the occupational data-based, automated outline procedure has been applied. From a validity standpoint, the most telling evidence is yet to come. Final judgments must wait until the statistical characteristics of the end products (the administered SKTs) are analyzed and, most importantly, judgment must be withheld until comments from SMEs on subsequent revisions of ATO-developed SKTs and from supervisor and co-worker judgments of SKT examinees' job knowledge can be assessed. Examinee comments will possibly be available from two sources: a brief survey administered to the examinee before he/she leaves the testing room and complaint letters sent to the USAF Occupational Measurement Center (USAFOMC). The expectation is that there will be significantly fewer comments than in the past regarding lack of job relatedness of test items.