Criterion Referenced Assessment: Establishing Content Validity of Complex Skills Related to Specific Tasks

David MacQuarrie, Ph.D.
Western Michigan University

Brooks Applegate, Ph.D.
Western Michigan University

Warren Lacefield, Ph.D.
Western Michigan University

ABSTRACT

Career and Technical Education (CTE) is a nationwide program that emphasizes training for primary, secondary, and post secondary educational stages for the career and workforce needs of today and tomorrow's society. Mandated indicators of success have been set in place and secondary schools are expected to improve student's skill levels in preparation for their next stage of education or employment. This study examines ways to measure proficiency in Automotive Service Technology (AST) skill ability domain levels, which consist of knowledge, concepts, and skills. The second part of the study examines the reliability and validity of an assessment method that is aligned with the AST foundational skills and ability levels needed by students or future employees and are intended to be a means to evaluate their readiness for their next educational stage.

CRITERION REFERENCED ASSESSMENT: ESTABLISHING CONTENT VALIDITY OF COMPLEX SKILLS RELATED TO SPECIFIC TASKS

The United States Department of Education's Strategic Plan for 2007 to 2012 outlines focused initiatives for educational reform (2007). The third goal of the Strategic Plan centers on students' successful transition between secondary education, post secondary education, and the workforce. Career and Technical Education (CTE) is a nationwide program that emphasizes training for primary, secondary, and post secondary educational stages for the workforce needs of today and tomorrow's society. Indicators of success have been mandated and schools are expected to improve student's skill levels to prepare the student for the next stage.

The context for the assessment validation and evaluation process was the CTE area of Automotive Service Technology (AST). State mandates for funding frequently require an AST program to meet the National Institute of Automotive Service Excellence (ASE) (2005) program certification standards. ASE is most noted for Automotive Technician (AT) competency certification, but also certifies AST training programs through the National Automotive Technician Education Foundation (NATEF) (2005). NATEF sets quality standards for AST programs that include current industry task listings, required tool and equipment lists, and a general description of skills that are assumed to be taught and learned. Unfortunately, two skill areas are apparently prerequisite, but are not clearly defined or assessed by NATEF or ASE and include Basic Vehicle Interval Maintenance Skills and Basic Vehicle Repair Skills, which form the Automotive Service Technology Foundational Skills (ASTFS) set.

Current reviews of certification or standardized assessment literature do not reveal a singular assessment that can measure ASTFS skills for secondary or post secondary education.

Neither can current assessments be identified for employers to use during employee screening, hiring, and training.

Recent efforts have been made to derive an accurate, specific, and valid listing of the automotive ―assumed" ASTFS skills and other skill domains that are hierarchically prerequisite to the AST tasks. A qualitative analytical process was used to identify specific ASTFS tasks and the underlying knowledge, concepts, and skills. Further, the assumed or underlying prerequisite skills were represented in such a manner that they could be taught, learned, and assessed (MacQuarie, Applegate, & Lacefield, 2008). The purpose of the ASTFSP Assessment is to provide a current or prospective AT employee with their proficiency level of the ASTFS as compared to industry criterion levels. Assessment procedures were aligned in accordance with the credentialing standards for educational and psychological testing (Joint Committee on American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education [Joint Committee], 1999). This paper highlights the second phase of this study which examines the reliability and validity of a criterion referenced proficiency assessment created for the ASTFS.

The assessment design, creation, and validation process that is described in this paper utilizes widely accepted measurement standards for a credentialing assessment (Joint Committee, 1999) in so far as the skills assessed are prerequisite in a hierarchy of learning the job level tasks performed in an employment setting. For instance, it is necessary to know the proper tools to use as well as understand safe operational practices and limitations for the use of those tools in order to cut off a stripped bolt or nut. Similarly, the measurement standards for educational testing and assessment (Joint Committee, 1999) were additionally referenced during the development process to ensure the usability of the assessment for multiple purposes.

REVIEW OF THE LITERATURE

The AST repair industry sector is expected to increase by 30.7% between the years 2004 and 2014 (US Department of Labor, 2007). CTE training for AST students exists at secondary and post secondary levels. Specific training for a manufacturer‘s line of vehicles begins shortly after employment if the automotive technician (AT) is evaluated as being ready. Thus, it is important that CTE properly service the educational needs of prospective AT employees. It is also important that AST employers can evaluate a prospective AT‘s technical proficiency levels quickly and accurately, prior to hiring or in order to evaluate a person‘s readiness for job level responsibilities, activities, and further training.

There are basically two assessment design routes available when assessing a person‘s ASTFS levels. The first route is a performance testing strategy of the ASTFS (Gronlund, 1998). This type of test typically relies on direct observation and judgement rather than traditional paper and pencil forced choice strategies (Worthen, White, Fan, & Sudweeks, 1999). Performance testing includes a simulation of the actual tasks, during which the testing candidate is observed and rated by an evaluator (Gronlund, 1998). An advantage is that this process can be very valid. Disadvantages include high expense, time consuming processes, individual administration requirements, need for evaluator training, and inter-rater reliability estimation techniques (Worthen, et al., 1999; Gronlund, 1998). Inter-rater variability in particular is difficult to solve and often results in low reliability and subjective results. Performance testing is commonly attempted in most of the AST training programs and is the preferred method of CTE teachers and is also a requirement of NATEF AST program certification (NATEF, 2005).

A second type of assessment strategy would include a norm-referenced test (NRT) or criterion-referenced test (CRT) to serve as an objective, psychometric measure of aptitude or proficiency (Worthen, et al., 1999). Psychometric testing can be defined as the science of measuring psychological aspects of a person such as knowledge, skills, abilities, or personality. Psychometric testing traditionally utilizes paper and pencil or computer adapted assessment techniques (Labor Law Center, 2005). An aptitude test is generally a broadly constructed measure of the cumulative effects of past learning (on a test taker) in order to predict future learning potential (Worthen, et al., 1999; Hopkins, 1998). Proficiency tests tend to be much more specific measures of a persons' ability to perform specific tasks at a specific criterion level, which would consistently parallel those levels typically found in authentic performance contexts.

Reliability estimates and empirical validity indicators require the application of various statistical procedures based on the type of assessment being evaluated (Joint committee, 1999). Reliability estimates for a norm-referenced, standardized achievement test do not work as well for CRT proficiency assessments due to the restriction of item homogeneity responses in the target population (Crocker & Algina, 1986). Although Cronbach's Coefficient Alpha is a versatile and widely used method for estimating the internal consistency of a typical scale or NRT (Worthen, et al., 1999), it simply is not as meaningful or useful in the context of CRT proficiency assessment (Crocker & Algina, 1986). A more meaningful reliability estimate is the test-retest correlation coefficient between two sets of test takers' scores on an assessment that is administered about two weeks apart in time. A better way to estimate reliability for a standardized proficiency assessment is to report the probability of a mastery decision using the same cut score or criterion on a parallel test. Subkoviak's (1976) Coefficient of Agreement (CGA) estimate for a mastery decision reports the probability that the test takers would be assigned mastery on a parallel test to the first test, based on results from a single test administration.

Validity can be defined as the degree to which concurrent evidence supports assessment scores (Joint committee, 1999). In the case of the ASTFS Proficiency Assessment (ASTFSP), the resulting scores are ideally the predictor measure and the ASTFS is the actual criterion level of the test taker. However, as the ASTFS construct is still newly delineated (MacQuarrie, Applegate, Lacefield, 2008), concomitant measures of this construct were not available to serve as validating correlates. Therefore, indirect correlates and contrasted group procedures were used to validate the ASTFSP Assessment.

In many cases the goal is to determine a person's readiness level for the next educational stage. In these situations the test taker's score should be compared to a criterion level and not to a population or similar general population alone. Therefore, it was decided that the purpose of the ASTFSP Assessment is to provide a current or prospective AT employee with their proficiency level of the ASTFS as compared to industry criterion levels. The best way to ensure that ASTFSP Assessment scores are meaningful is to associate them with industry ASTFS criterion cut-score levels, such as those for certification purposes.

Bob Clark, a technical specialist in the Special Testing Programs for ASE was the resource sought for expert guidance on standard cut score procedures (B. Clark, personal communication, September 2, 2005). Clark indicated that ASE cut scores were set for each scale area assessment using a "modified Angoff procedure," which he claimed is common for high stakes tests (2005). Clark then asserted that items were carried forward for future tests using "pre-equating" procedures based on Item Response Theory (IRT) techniques. In addition to reporting the single cut-score result of pass / fail for test takers, ASE also reported the number of correct responses for each section of the test.

SKILL DEFINITION AND REPRESENTATION

The ASTFS skills and tasks and were defined and delineated in a previous paper titled, *Criterion Referenced Assessment: Delineating Curricular Related Performance Skills Necessary for the Development of a Table of Test Specifications*," (MacQuarrie, Applegate, & Lacefield, 2008). A summary of the findings from that previous paper will be reported here to maintain continuity of the material. The ASTFS are listed in two scales: the Basic Vehicle Repair Skills (BVRS) and the Basic Vehicle Interval Maintenance Skills (BVIMS).

General categorical listings of tasks often are ambiguous in the sense that they fail to further delineate lower levels of prerequisite skills. However, they do have utility for illustrating skill and task hierarchies. Refer to Table 1 and 2 for a listing of the Table of Test Specifications for each of the two scales of the ASTFS. Within sub-categories, the general units, tasks, and objectives used for further defining the skills are delineated.

Procedures were followed in accordance with credentialing assessment standards and published best practices to create the ASTFSP Assessment. The ASTFSP Assessment is a criterion referenced mechanical aptitude assessment of multiple choice design. Several iterations of preliminary and pilot studies using both qualitative and quantitative processes provided information to improve item quality concerning reliability and validity.

Table 1.

Basic Vehicle Repair Skills Table of Test Specifications for the ASTFS

	Skill Levels			
Basic Vehicle Repair Skills Sub-Scale Categories	Knowledge Level	Comprehension Level	Application / Analysis Level	Percentage
Oxy-Acetylene Torch Safe Usage Scale				
Oxy-Acetylene Torch Set-up	21, 25	22		12.50%
Oxy-Acetylene Torch Storage		26		4.17%
Oxy-Acetylene Torch Practices		23	24	8.33%
Sub-Scale Percentage of Test Scale Percentage ()	8% (33%)	13% (50%)	4% (17%)	26%
Mechanical Aptitude & Safe Tool Use Scale				
Pneumatic tools and equipment	27	28		8.33%
Electrical power tools and equipment	29		30	8.33%
Hand tool selection and use	31	32		8.33%
Mechanical Aptitude			33, 34, 35, 36	16.67%
Sub-Scale Percentage of Test Scale Percentage ()	13% (27%)	8% (18%)	21% (45%)	44%
Facility Equipment Use and Safety Scale				
Hoists and jack use		37	38	8.33%
Fire extinguisher selection and use	39			4.17%
Ventilation		40		4.17%
Personal Protective Equipment	41, 42			8.33%
Environmental Concerns	43, 44			8.33%

Sub-Scale Percentage of Test Scale Percentage ()	21% (63%)	8% (25%)	4% (13%)	30%
Percentage of Test	41.7%	29.2%	29.2%	100.0%

Table 2.

Basic Vehicle Interval Maintenance Skills Table of Test Specifications for the ASTFS

	Skill Levels			
Interval Maintenance Sub-Scale Categories	Knowledge Level	Comprehension Level	Application / Analysis Level	Scale Total Percentage
3,000 to 7,500 mile Interval Maintenance Sub-Scale:				
Change oil and filter		3, 5		10%
Lube chassis and drive-train				
Check/Service all fluid levels	2	7	17	15%
Check/Locate Fluid leaks		14, 16		10%
Lube vehicle access features				
Check/Service clutch free play				
Check/Service drive belts		1		5%
Perform Safety Inspection				
Check/Service tire pressure	4			5%
Check/Service all hoses				
Check/Service battery and cables	13			5%
Check/Service MIL light, engine, body codes			20	5%
Sub-Scale & Total Scale Percentage ()	27% (15%)	54% (30%)	18% (10%)	55%
One year or 15,000 mile Interval Maintenance Sub-Scale:				
All of the 3,000 mile maintenance areas:				
Check/Service tires and wheels	12			5%
Replace air filter				
Check/Service all hoses (Coolant and Vacuum)				
Check/Service cooling system & A/F protection		8, 9		10%
Clean radiator externally			11	5%
Check/Service tires & wheels (rotate tires/wheels)		18	19	10%
Check/Service emissions filter				
Check/Service brake components				
Check/Service steering and suspension components				
Check/Service vehicle condition (cosmetically)				
Maintenance the battery (if applicable)				
Check/Service C.V. joints and suspension				
Lube CV joint boots				
Lube door seals				

Replace spark plugs (optional)				
Replace air cabin filter				
Sub-Scale & Total Scale Percentage ()	16% (5%)	50% (15%)	33% (10%)	30%
Two year or 30,000 mile Interval Maintenance Sub-Scale:				
All of the items of the 15 K maintenance				
Flush brake fluid		6		5%
Flush auto-trans fluid (optional)				
Flush cooling system		10		5%
Replace fuel filter (optional)			15	5%
Sub-Scale & Total Scale Percentage ()	0% (0%)	66% (10%)	33% (15%)	(15%)
Total Skill Level Percentage of the Test	20%	55%	25%	100%
Note a Refer to scale. b Dependant on specific certification. c Dependant on repair facility option selection.				

PURPOSE OF THE STUDY

This second phase of the study examines and evaluates the reliability and validity of an assessment that is aligned with the ASTFS needed by students or future employees that is intended to evaluate their readiness for their next educational stage. This paper describes assessment design processes and not research processes and will, therefore be presented in manner that fits typical assessment procedures. In order to achieve the purpose for this study, objectives that align with typical assessment design and construction processes are used.

1. The assessment design and construction processes utilized an assessment purpose and methods to ensure both the content and ability domains were proportionately aligned with a highly recognized content or skill area.
2. The assessment‘s item writing processes referenced and reflected the content and ability domains and were then improved through preliminary item try-outs and pilot studies.
3. Assessment and item reliability processes included empirical evidence of reliability estimations from item analyses procedures through both internal consistency and external comparison estimations.
4. Assessment and item validation processes included empirical evidences of convergent and/or discriminate validity.

The first objective was previously completed and described in detail in a previous paper, as explained in the previous section of this paper. The processes of this phase included: internal and external reliability estimations, content validation, and convergent and discriminate validation procedures including: contrasted group methods, concurrent correlations, and IRT analytical procedures. Initial standardization procedures were implemented using an objective means of deriving the passing cut scores.

METHODS

The Table of Test Specifications (ToTS) was fulfilled by specific items being written

using the research resources used to complete the ASTFS list. Multiple techniques were employed to estimate the psychometric properties of the ASTFSP Assessment. A preliminary study of the ASTFSP Assessment included procedures for empirical validation focused on improving the ASTFSP Assessment and gathering evidence for construct validity (MacQuarrie, 2005). This study expanded the participant groups to include CTE high school students and further statistical analyses. In addition, IRT procedures were employed to assess the utility of the ASTFSP Assessment items and scores for the intended goal of measuring ASTFS ability levels.

INSTRUMENTATION: The ASTFSP Assessment used a four choice multiple choice format for items forming two scales. The first scale included items related to the BVIMS area, such as: lubrication replacement procedures, coolant selection, and tire pressure checking and correcting procedures. The second scale included items related to the BVRS area, such as: hand and power tool safety, selection, and procedures, fastener selection and uses, and oxygen-acetylene torch safety procedures.

PROCEDURES: The ASTFSP Assessment items were specifically written to fit the ASTFS ToTS plan proportions of content and ability domain as displayed in Tables 1 & 2. The item writing process was performed by the primary author of this paper using research references from the previous delineation process as well as industry related case studies in a manner as described in typical item writing texts (Worthen, White, Fan, & Sudweeks, 1999; Gronlund, 1998; Hopkins, 1998).

The empirical construct validation process required a two part approach because of the lack of concomitant measures for the ASTFS construct. The two approaches for gathering convergent and discriminate evidence included both: contrasted groups and correlating the ASTFSP Assessment scores with criterion measures. This study used two primary groups of participants with opposing levels of AST skills for the administration and gathering of the ASTFSP Assessment data. They first group was the AST experienced group and was composed of three subgroups: 1) AST experts working in industry 2) AST teacher members of the Automotive Youth Educational Systems (AYES) (2005) and 3) a group of AST high school students age 16 to 20 years old near the end of a year‘s training. The second group consisted of non-AST experienced participants in two subgroups: 1) non-AST high school students age 16 to 20 years old, and 2) non-AST teachers. A survey question within the ASTFSP Assessment identified a sixth potential group, from the first two subgroups within the two primary groups, who self-identified themselves as AST hobbyists and who would likely vary widely in AST skill level.

Reliability estimates for internal consistency, test-retest correlation, and Subkoviak‘s (1976) CGA estimate for mastery tests are presented below. Descriptive statistics for the test takers are reported by group and subgroup. Discriminate validation included MANOVA procedures to test the hypothesis of mean differences on the dependant variables, BVIMS and BVRS, among various groups: AST experts, AST teachers, Non-AST students, non-AST teachers, AST students, and AST hobbyists. A canonical discriminant function analysis (DFA) was conducted to determine whether ASTFSP Assessment scores could be used to differentiate between the six AST groups, but is not reported here to reduce redundancy.

The convergent validation approach included correlating the ASTFSP Assessment scores with criterion measures of AT developmental indicators. Developmental indicators were reported by an AST test taker‘s supervisor on a second performance rating scale and survey completed while the participating technician was completing the ASTFSP Assessment. Correlations are reported between the ASTFSP Assessment scores and developmental indicators such as ASE

certifications, State of Michigan Certifications, and work duty responsibilities.

The expert experiential group's scores were then used to establish cut-scores in an objective manner. The method used is similar to a contrasted groups approach separating the expert's scores from those of other groups. The current study also allowed IRT procedures to be used to further evaluate the ASTFSP Assessment items and to measure AST trait levels of ability. The first two expert subgroups' ASTFSP Assessment data were used to calculate the item difficulty for each item. Item difficulties were used in a manner similar to the way the Angoff procedure uses experts. Traditional Angoff procedures typically use selected experts to directly estimate the probability of a mythical minimally competent person would get correct for each item (Standard, 2008). The Angoff procedure would be offset for this assessment to ensure objectivity by way of contrasted groups. The offset would be completed by summing the selected experts' actual item difficulties together for the two expert sub-sample groups, thereby deriving a set of appropriate cut scores based on actual experience.

RESULTS

The first and previous part of this study set the stage for construct validity of the ASTFSP Assessment with the ToTs. The second objective was:

> *The assessment's item writing processes referenced and reflected the content and ability domains and were then improved through preliminary item try-outs and pilot studies.*

To complete the second objective the ASTFSP Assessment items were specifically written to fit the ASTFS ToTS plan proportions of content and ability domain as displayed in Tables 1 & 2. The item writing process was performed by the primary author of this paper using research references from the previous delineation process as well as industry related case studies gathered from industry personnel. Preliminary item try-outs and pilot studies were used to improve the items in multiple ways. First, qualitative feedback was gathered for the items as experts completed the assessment along with an instrument rubric. Second, iterative improvements used simple Item Analysis procedures for monitoring: response proportions, item completion, and proportions correct. Each administration resulted in an iteration of assessment improvement as well as provides a deeper and more objective result due to an increased number of participants.

The third objective was initiated during the previous section with simple Item Analyses. The third objective involving reliability is performed on the data that is gathered from the validity study due to assessment design and is as follows:

> *Assessment and item reliability processes included empirical evidence of reliability estimations from item analyses procedures through both internal consistency and external comparison estimations.*

To complete the third and fourth objective related to reliability and validity a two part approach was used due to the lack of concomitant measures of the ASTFS construct. The first approach used contrasted groups and the second approach correlated ASTFSP Assessment scores to developmental indicators. Refer to the descriptive statistics in Table 3, which depict score and scale score means, number of participants in a group, and standard deviations that will need to be statistically tested for differences. Refer to Figure 1 for the number of ASE certifications

possessed by the AST Industry Expert Group.

Internal consistency reliability estimates for the 39 question version of the ASTFSP Assessment are reasonable for a proficiency assessment: (n = 354) overall scale α = .730 with a Confidence Interval estimate of $\alpha \geq .688 \leq .769$ (Barnette, 2005), BVIMS scale α = .602, and BVRS scale α = .590 (Worthen, White, Fan, & Sudweeks, 1999). However, Cronbach Alpha α is lower bound estimate of reliability when used for a criterion referenced assessment (Crocker & Algina, 1986).BVRS scale α = .590 (Worthen, White, Fan, & Sudweeks, 1999). However, Cronbach Alpha α is lower bound estimate of reliability when used for a criterion referenced assessment (Crocker & Algina, 1986).

External reliability compares assessment scores in time or with another parallel form of the test. A study of stability reliability estimates for the ASTFSP Assessment within a 20 day period indicated excellent results (n = 24) ρ = .908 with a Confidence Interval of $\alpha \geq .796 \leq .959$ (Barnette, 2005). Reliability estimates for the three cut scores set on a test-retest mastery decision is (n=24): .96, .88, and 1.00, which is interpreted as the proportions of participants that were assigned the same mastery score decision as 96%, 88%, and 100%, which is very good. Subkoviak's (1976) CGA estimate for a mastery decision for the ASTFSP Assessment (n=354) was good at .741, which is interpreted to mean that an individual would have a 74% lower bound probability that he or she would be assigned to the same mastery (or non-mastery) state on a second testing that was parallel to the first test. Refer to Figure 2 for a graphical representation of the Coefficient of Agreement for the ASTFSP Assessment. Stability (test-retest) reliability, when available, is a better indicator of reliability than internal reliability or even CGA estimation since the ASTFSP Assessment is a proficiency assessment (Crocker & Algina, 1986).

Figure 1. Number of ASE Certifications Possessed by AST Expert Group Members

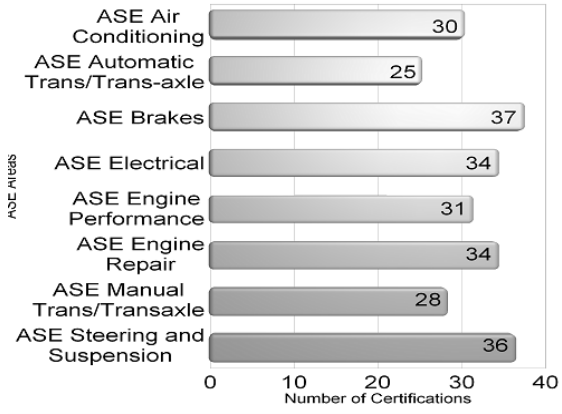


Figure 2. Graph of the Coefficient of Agreement for the ASTFSP Assessment

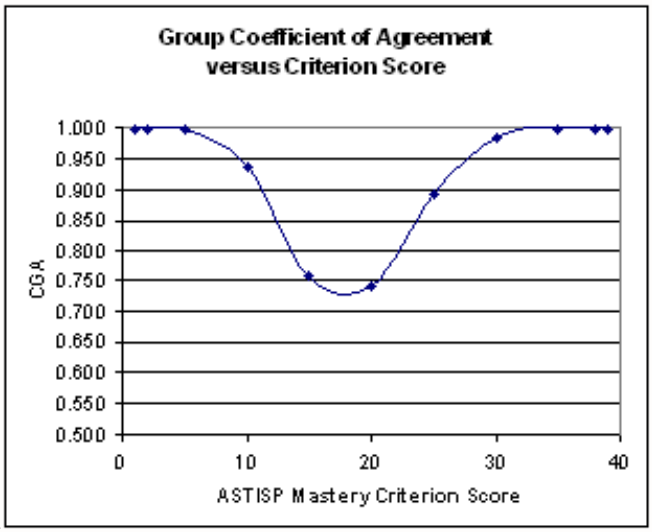


Table 3.

Descriptive Statistics for the Five Group's Scores

Experiential Test Group		Mean	N	Std. Deviation
Expert Industry	BVIMS Score	9.44	62	1.554
	BVRS Score	13.29	62	2.329
	ASTFSP Score	22.73	62	3.235
Expert Teacher	BVIMS Score	10.50	14	1.653
	BVRS Score	14.71	14	2.813
	ASTFSP Score	25.21	14	3.704
Non-Expert Student	BVIMS Score	3.92	39	1.628
	BVRS Score	6.85	39	1.913
	ASTFSP Score	10.77	39	2.518
Non-Expert Teacher	BVIMS Score	6.75	12	2.094
	BVRS Score	9.25	12	2.050
	ASTFSP Score	16.00	12	2.412
Hobbyist	BVIMS Score	5.70	10	2.584
	BVRS Score	8.70	10	3.234
	ASTFSP Score	14.40	10	4.881
AST High School Student	BVIMS Score	6.97	217	2.050
	BVRS Score	10.28	217	3.009
	ASTFSP Score	17.25	217	4.295
Total	BVIMS Score	7.16	354	2.506
	BVRS Score	10.53	354	3.359
	ASTFSP Score	17.69	354	5.252

CONTRASTED GROUPS VALIDATION RESULTS

The fourth objective summarizes the validity study design and uses the data gathered to make conclusions about the ASTFSP Assessment's meaning and not for purposes of the groups' differences, as in research. The fourth objective is as follows:

> *Assessment and item validation processes included empirical evidences of convergent and/or discriminate validity.*

There isn't another direct assessment parallel to ASTFSP Assessment that measures the ASTFS. Therefore, statistical discriminate differences were sought between natural groups based on ASTFS experience. MANOVA procedures were performed to test for differences between the dependant variable scale scores, BVIMS and BVRS, of the ASTFSP Assessment for the six groups: AST experts, AST teachers, Non-AST high school students, non-AST teachers, AST high school students, and AST hobbyists.

Box's test indicated the MANOVA equality assumption was violated, $p < .05$ and thus, required Dunnett's C correction for the post hoc analyses. Results of the MANOVA for mean differences on BVIMS and BVRS scales for the six groups were statistically significant *Wilks' Lambda* = .529, $F(10, 694) = 26.041$, $p < .001$ and partial $\eta^2 = .273$, indicating that 27% of the variance was accounted for in the model. Univariate analyses revealed both statistical and practical effects for each dependent variable, followed by interesting Dunnett's C post hoc pair-wise means comparisons among groups. Refer to Tables 4 and 5, respectively for ANOVA results and the post hoc analysis. Refer to Figures 3 and 4 for a graphical representation of the estimated means for each group and scale of the ASTFSP Assessment.

In summary, the MANOVA results indicate there are statistically significant differences between the two primary experiential groups and sub-groups for both scales of the ASTFSP Assessment: those with ASTFS experience and those without. Therefore, ASFTSP Assessment could allow a detection of discriminate differences, thus indicating a measure of validity for the assessment.

Table 4.

ANOVA Results for Each Predictor Variable

Source	*df*	*F*	*p*	η^2
BVIMS Scale Scores	(5, 348)	49.096	< .001	0.414
BVRS Scale Scores	(5, 348)	34.250	< .001	0.330

Table 5.

Results for the Post Hoc Analyses

Source	Group	Industry Expert	Expert Teacher	Non-AST Expert Student	Non-AST Expert Teacher	AST Hobbyist
BVIMS Scale Scores	Expert Teacher	$p > .05$	-			
	Non-AST	$p < .05$	$p < .05$	-		
	Non-AST	$p < .05$	$p < .05$	$p < .05$	-	
	AST Hobbyist	$p < .05$	$p < .05$	$p > .05$	$p > .05$	-
	AST HS Student	$p < .05$	$p < .05$	$p < .05$	$p > .05$	$p > .05$
BVRS Scale Scores	Expert Teacher	$p > .05$	-			
	Non-AST	$p < .05$	$p < .05$	-		
	Non-AST	$p < .05$	$p < .05$	$p < .05$	-	
	AST Hobbyist	$p < .05$	$p < .05$	$p > .05$	$p > .05$	
	AST HS Student	$p < .05$	$p < .05$	$p < .05$	$p > .05$	$p > .05$

Note: Dunnett's C post hoc analyses correction

Figure 3. Estimated Marginal Means for the BVIMS Scale for the Six Groups

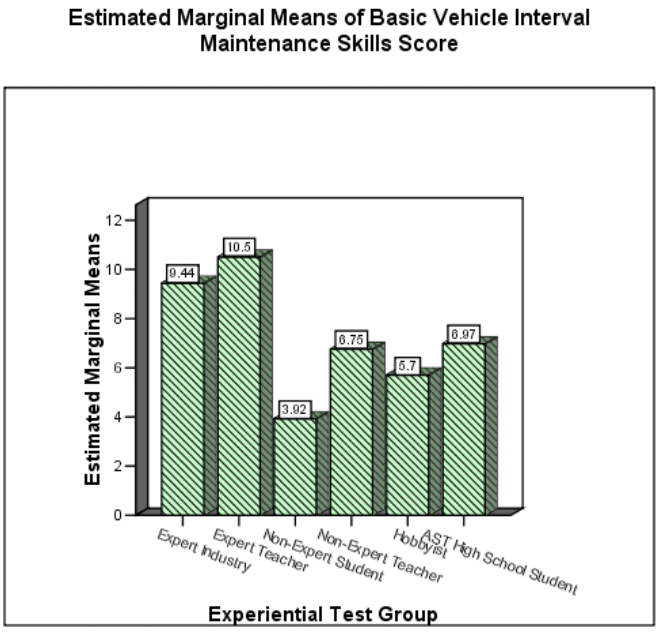
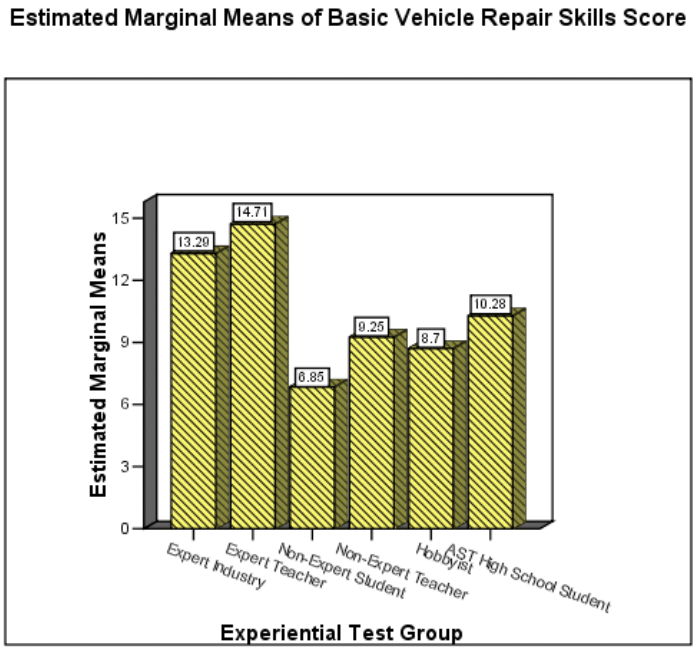


Figure 4. Estimated Marginal Means for the BVRS Scale for the Six Groups



DEVELOPMENTAL INDICATOR RELATIONSHIPS

To evaluate convergent validity relationships, additional data was gathered on the AST experts to test correlation relationships. While AST industry technicians were completing the ASTFSP Assessment, their supervisors completed a performance rating scale and survey on the participating technician. Moderate correlations existed between the ASTFSP Assessment scores and developmental indicators such as ASE certifications, State of Michigan Certifications, and work duty responsibilities. Work duty responsibilities is an anchor rating scale continuum on which a mark is assigned by the supervisor for each technician within his or employ, which range from ―dichanging mechanic" to ―top diagnostician of complex vehicle problems". The positive correlations among the various measures indicate the ASTFSP Assessment is measuring aspects of a person's work-related to performance instead of simply paper and pencil test taking abilities. Refer to Table 7 for a list of Spearman's Correlations showing a developmental progression with the ASTFSP Assessment.

Table 7.

Developmental Correlations for the ASTFSP Assessment

	ASTFSP Scores	Work Duty Responsibilities	Number of State of Michigan Certifications	Number of ASE Certifications
ASTFSP Scores	1.000			
Work Duty Responsibilities	0.482	1.000		
Number of State of Michigan Certifications	0.444	0.693	1.000	
Number of ASE Certifications	0.368	0.668	0.864	1.000

SCALE SETTING OF CUT SCORES

Finally, two cut-scores were derived using an objective approach based on contrasted groups to separate the subjects into three grouping categories: "Non-Expert", "Minimal Knowledge", and "Minimally Competent" level groups. The cut score results were then evaluated using DFA classification procedures between the cut score groups. Refer to Table 8 for the results of the cut score DFA classification results.

Table 8.

DFA Classification Results for the Cut Score Groups

		Predicted Group Membership			
	Passing Cut Scores	Non-AST Experts	Minimal Knowledge Level	Minimally Competent	Total
Original Count	Non-AST Experts	132	0	0	132
	Minimal Knowledge Level	0	194	0	194
	Minimally Competent	0	0	28	28
Percentage	Non-AST Experts	100.0	.0	.0	100.0
	Minimal Knowledge Level	.0	100.0	.0	100.0
	Minimally Competent	.0	.0	100.0	100.0

100.0% of original grouped cases correctly classified.

The "Minimally Competent" level spanned a score percentage range from approximately 65% to 79%. There were a limited number of participant scores near 80% therefore, the "Competent" cut score level is reserved for scoring future ASTs who have been purposefully and effectively trained in the ASTFS.

The first part of this paper has presented validity and performance results for various expert levels in the field of AST, teachers, and ATs. The second part of this report will present additional results from a high school CTE ASTFSP Assessment. The larger sample of CTE high school students allowed IRT procedures to be used to further evaluate the validity of the ASTFSP Assessment items and ASTFS ability levels.

LATENT TRAIT DIMENSIONALITY

Content validity is important, but more important is the validity of the latent trait: the ability domain. To analyze the validity of the ability variances of the ASTFSP Assessment items IRT procedures were performed. IRT procedures allow the plotting of ICC's along the latent trait continuum to gain insight into each item's functionality of performance being measured by the assessment. BILOG MG was used to estimate a one parameter IRT on the ASTFSP Assessment data. Refer to Figures 5 and 6 for a graphic depiction of the Total Information and Standard Error for each scale of the ASTFSP Assessment. The high level of Standard Error depicted in the graph is attributed to the AST high school group, as can be verified by the Standard Deviations reported in Table 3.

Item level parameters for the BVIMS scale indicate that the items are discriminating and

vary across the Latent Trait ability level. Chi-Square item fit statistics for the BVIMS scale indicate the items 1, 2, 4, 5, 7, 13, and 15 may be better fit by a higher parameter model. Item level parameters for the BVRS scale indicate that the items are discriminating and vary across the Latent Trait ability level. Chi-Square item fit statistics indicate that items 16, 18, 19, 21, 22, 24, 26, 35, and 39 may be better fit by a higher parameter model. As additional data are collected the two parameter IRT model will be estimated and evaluated. Refer to Tables 13 and 14 for Item Parameter Statistics. Refer to Figure 7 for a graphic representation of the ICC's for the BVIMS and BVRS scales as indicated by the dark and lighter shades, respectively.

Table 13.

BVIMS Scale IRT Item Parameters

Item		Threshold	χ^2	*p*	Df
BVIMS01	Parameters	-1.917	27.6	0.001	6.0
	S.E.	0.251			
BVIMS02	Parameters	-1.448	24	0.001	5.0
	S.E.	0.23			
BVIMS03	Parameters	-1.483	10.2	0.116	6.0
	S.E.	0.219			
BVIMS04	Parameters	-1.141	30.4	0.001	6.0
	S.E.	0.204			
BVIMS05	Parameters	-0.675	24	0.001	6.0
	S.E.	0.189			
BVIMS06	Parameters	-0.809	9.1	0.170	6.0
	S.E.	0.19			
BVIMS07	Parameters	0.28	27.2	0.001	5.0
	S.E.	0.189			
BVIMS08	Parameters	0.114	8.5	0.288	7.0
	S.E.	0.178			
BVIMS09	Parameters	0.215	4.8	0.779	8.0
	S.E.	0.181			
BVIMS10	Parameters	0.695	8.9	0.261	7.0
	S.E.	0.191			
BVIMS11	Parameters	0.723	2.5	0.870	6.0
	S.E.	0.188			
BVIMS12	Parameters	0.96	6.5	0.480	7.0
	S.E.	0.196			
BVIMS13	Parameters	1.239	27.8	0.001	7.0
	S.E.	0.216			
BVIMS14	Parameters	1.239	13.5	0.061	7.0
	S.E.	0.207			
BVIMS15	Parameters	2.008	16.9	0.018	7.0

	S.E.	0.262			

Table 14.

BVRS Scale IRT Item Parameters

Item		Threshold	χ^2	*p*	Df
BVRS16	Parameters	2.612	26.800	0.001	5.0
	S.E.	0.338			
BVRS17	Parameters	2.175	12.300	0.056	6.0
	S.E.	0.286			
BVRS18	Parameters	1.598	25.000	0.001	6.0
	S.E.	0.248			
BVRS19	Parameters	1.033	13.800	0.017	5.0
	S.E.	0.225			
BVRS20	Parameters	0.511	10.200	0.117	6.0
	S.E.	0.211			
BVRS21	Parameters	0.448	23.000	0.001	6.0
	S.E.	0.216			
BVRS22	Parameters	0.174	27.000	0.001	7.0
	S.E.	0.213			
BVRS23	Parameters	0.486	2.800	0.947	8.0
	S.E.	0.205			
BVRS24	Parameters	0.074	21.400	0.002	6.0
	S.E.	0.216			
BVRS25	Parameters	0.137	5.100	0.651	7.0
	S.E.	0.207			
BVRS26	Parameters	0.259	23.800	0.001	7.0
	S.E.	0.22			
BVRS27	Parameters	0.162	5.500	0.708	8.0
	S.E.	0.207			
BVRS28	Parameters	0.014	6.100	0.523	7.0
	S.E.	0.207			
BVRS29	Parameters	0.181	2.700	0.908	7.0
	S.E.	0.207			
BVRS30	Parameters	0.339	11.300	0.127	7.0
	S.E.	0.218			
BVRS31	Parameters	0.339	10.400	0.166	7.0
	S.E.	0.217			
BVRS32	Parameters	0.272	10.2	0.180	7.0
	S.E.	0.208			
BVRS33	Parameters	0.798	8.2	0.314	7.0
	S.E.	0.233			
BVRS34	Parameters	0.632	12.2	0.095	7.0
	S.E.	0.214			
BVRS35	Parameters	0.845	25.8	0.001	7.0
	S.E.	0.22			
BVRS36	Parameters	1.206	9.9	0.193	7.0
	S.E.	0.256			
BVRS37	Parameters	1.097	4.5	0.726	7.0
	S.E.	0.242			

BVRS38	Parameters	1.729	2.7	0.840	6.0
	S.E.	0.296			
BVRS39	Parameters	1.605	19.9	0.006	7.0
	S.E.	0.275			

Figure 5. BVIMS Scale Information and Standard Error

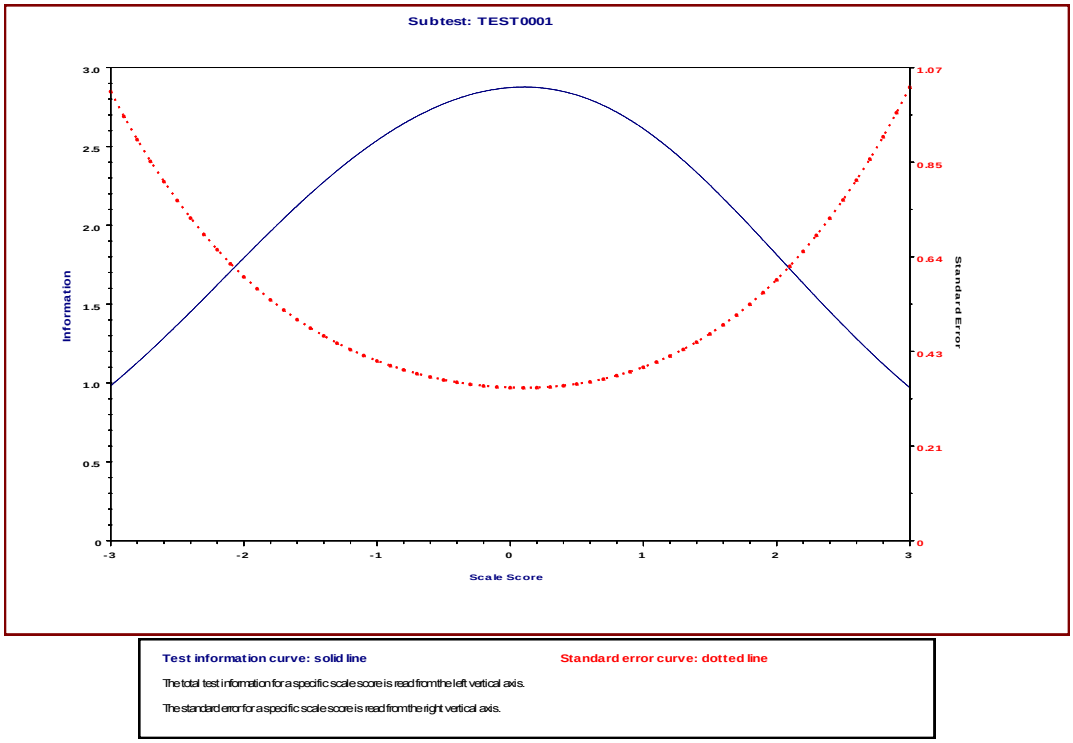
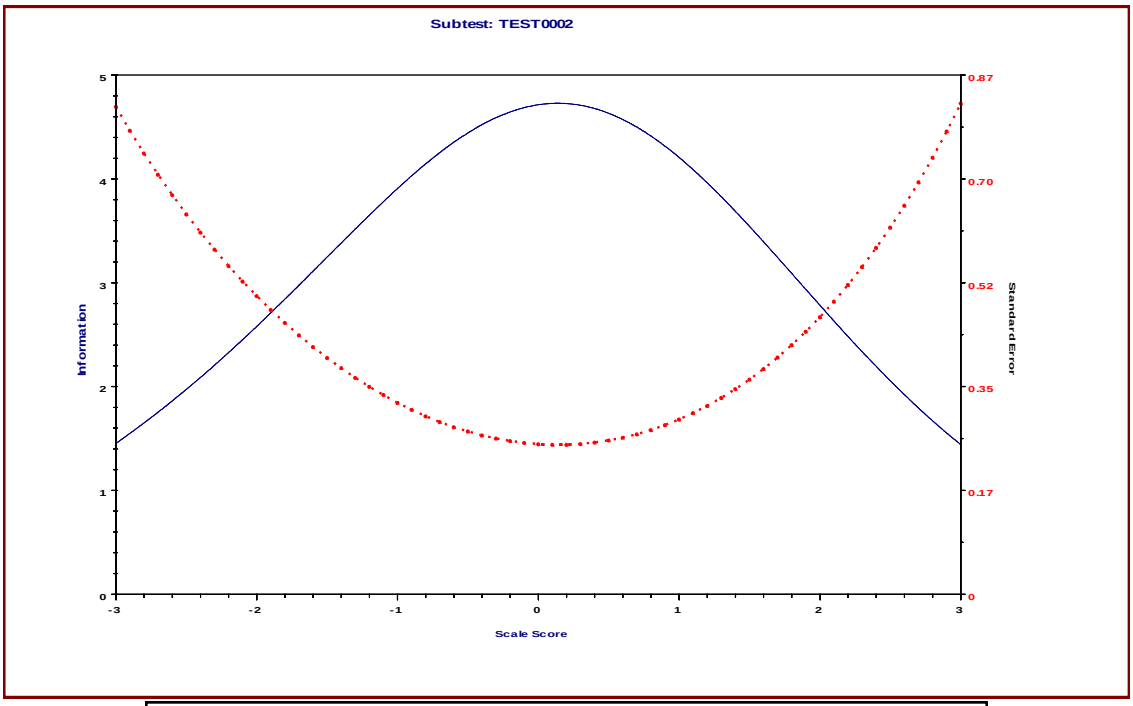


Figure 6. BVRS Scale Information and Standard Error

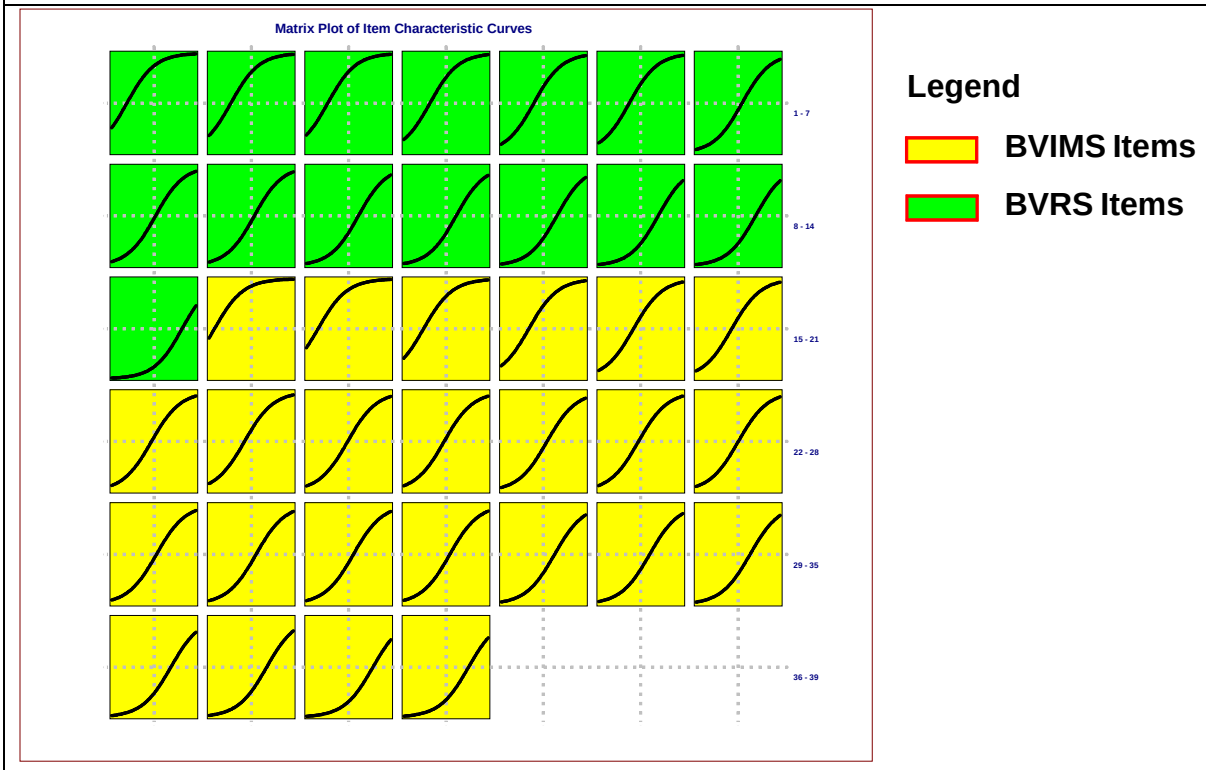


Test information curve: solid line **Standard error curve: dotted line**

The total test information for a specific scale score is read from the left vertical axis.

The standard error for a specific scale score is read from the right vertical axis.

Figure 7. *Item Characteristic Curves for the ASTFSP Assessment by Scales*



In summary, the IRT statistics indicate that the ASTFSP Assessment is measuring a varied ability level for each of the two scales. Further IRT analysis would be beneficial for a two and three parameter model for all items of the ASTFSP Assessment in future administrations.

DISCUSSION, IMPLICATIONS, AND OPPORTUNITIES

The ASTFSP Assessment is useful as a criterion referenced mechanical aptitude assessment for making group or individual level decisions. Several important points have emerged while viewing the ASTFSP Assessment results. The first point is that an AST technician should possess a "Minimally Competent" or higher level of the ASTFS as it would benefit their customers, their employers, and themselves. Additionally, an educational organization should consider including the direct instruction and assessment of the ASTFS due to the positive correlation between an AST technician's ASTFSP Assessment score and their success concerning the obtainment of AST developmental indicators such as higher work duties and the obtainment of certifications.

The second point focuses on the average ASTFSP Assessment scores obtained by AST industry experts of 58% and the average AST teacher score of 64%. These score levels are lower than expected for practicing AST technicians and teachers and is attributed to both a current and past lack of professional development specific to the ASTFS. There is a potential for growth among AST industry experts, AST teachers, and CTE high school level AST students.

The ASTFSP Assessment could be administered to multiple groups for various reasons.

Secondary level CTE students can be administered the ASTFSP Assessment to indicate readiness to enter the career field or post secondary level CTE. Post secondary level CTE students can be administered the ASTFSP Assessment to indicate readiness to enter the career field. ASTFSP Assessment scores from practicing or prospective AT's can indicate a need for professional development in the BVIMS or BVRS scale areas.

In closing, the discovery and development of both the ASTFS and the ASTFSP Assessment can assist schools, AST programs, and employers in evaluating the development level of AST's or AST students. AST students would likely benefit from effectively learning the ASTFS most if they were to learn them prior to the NATEF task lists as they are underlying skills. It would seem ideal for high school level students to possess a higher level of ASTFS proficiency to enable a student to learn and perform more effectively at the system level of AST duties and tasks. Additionally, all transportation technicians would most likely need most if not all of the ASTFS and may be transferable to other transportation and industrial areas. Future plans for further evaluating the potential of the ASTFSP Assessment is currently in planning stages for a predictive validity study and for AT's, AST students, and other technicians. Interested organizations or parties are invited to inquire, volunteer assistance, or support.

REFERENCES

Automotive Service Excellence. (2003). *About ASE*. Retrieved December 6, 2003 from http://www.asecert.org/subchannels/about_profile.cfm

Automotive Youth Educational Systems. (2005). *Automotive youth educational systems: The passport to a rewarding automotive career.* Retrieved March 20, 2005 from http://www.ayes.org/index.asp

Barnette, J. J., (2005). ScoreRel CI: Software for computation of commonly used score reliability estimates. *Educational and Psychological Measurement, 65*, 980-983

Crocker, L., Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont CA: Wadsworth Group/Thompson Learning.

Gronlund, N. E. (1998). *Assessment of student achievement (6th ed.).* Needham Height, MA: Allyn & Bacon.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation (8th ed.).* Needham Heights, MA: Allyn & Bacon.

Joint Committee on American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association

Labor Law Center. (2005). *Dictionary.laborlaw.com*. Retrieved April 20, 2005 from http://encyclopedia.laborlawtalk.com/Psychometrics

MacQuarrie, D.L., Applegate, B., Lacefield, W. (2008). *Criterion Referenced Assessment: Delineating Curricular Related Performance Skills Necessary for the Development of a Table of Test Specifications.* Journal of Career and Technical Education, Current Publications

MacQuarrie, D.L. (December, 2005). *Automotive service technology intersectional skills proficiency assessment*. UMI Dissertation Services, 129. (UMI No. 3197563)

National Automotive Technicians Education Foundation. (2005). *About us.* Retrieved December 29, 2004 from http://www.natef.org/about.cfm

Subkoviak, M. J. (1976). Estimating the reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13(4).* 265-276

U.S. Department of Education. (February, 2007). *Strategic plan for fiscal years 2007–12.*

Washington, DC: Retrieved February 10, 2007 from http://www.ed.gov/about/reports/strat/plan2002-07/index.html

U.S. Department of Labor Bureau of Labor Statistics. *(2007). Occupational Outlook Handbook 2006-2007 Edition: Tomorrow's jobs.* Retrieved February 14, 2007 from http://www.bls.gov/oco/oco2003.htm

Standard-setting study. (2008, November 4). *Wikipedia, The Free Encyclopedia*. Retrieved 22:27, November 10, 2008, from http://en.wikipedia.org/w/index.php?title=Standard-setting_study&oldid=249653158

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R., (1999). *Measurement and assessment in schools (2nd ed.).* Reading, MA: Addison Wesley Longman.

Whittington, M. S., & Raven, M. (1995). Learning styles: An assessment - An application. *NACTA Journal, 39*, 6-10.