Adaptive Normalization of Assessment Scores: A Multi-Study Validation Approach

Jeremiah T. Folsom-Kovarik, Angela Woods,
Daniel Wilson, Joseph Cohn
Soar Technology, an Accelint company
Orlando, FL
Jeremiah@soartech.com, Angela.Woods@soartech.com,
Daniel.Wilson@soartech.com, Joseph.Cohn@soartech.com

Lee Sciarini, Beth Atkinson
Naval Air Warfare Center
Training Systems Division
Orlando, FL
Lee.W.Sciarini.mil@us.navy.mil,
Beth.F.Atkinson.civ@us.navy.mil

ABSTRACT

Training and education programs often employ criterion-referenced evaluation, where individuals must perform tasks to a defined standard under specified conditions. However, domains that require high-performance differentiation, selection, and career progression also depend on norm-referenced comparisons, evaluating individuals relative to their current and recent past peers.

This paper describes a series of studies culminating in a statistical and artificial intelligence (AI)-driven model that continuously normalizes scores, enabling comparison in dynamic environments where assessment standards, grading criteria, and course structures evolve. The model estimates the similarity of assessment components drawn from different time frames to prevent comparisons of dissimilar data, ensuring that only compatible measures are analyzed. Additionally, it is robust to missing data caused by course changes. This model mitigates the impact of curriculum changes on rolling norm groups, preserving the integrity of comparative evaluation.

Initial studies defined the challenge of comparing individuals with a rolling norm group in a high-performing military population. Subsequent work identified which metrics are effective to measure score similarity, relate assessments across course versions, and make comparisons more robust to missing data. Final studies empirically validated the approach by successfully predicting historical outcomes using the normed model, demonstrating the model makes effective comparisons possible when there are planned and unplanned variations in course and assessment structure. Collectively, these studies provide support that the enhanced model improves accuracy and provides actionable insights about the data available for comparing performances.

By improving the accuracy of norm-referenced comparisons, this approach is applicable in numerous teaching and training contexts to support course syllabus adjustments, enhance individual development, and optimize allocation of resources. More accurate information for selection and evaluation processes can contribute to improved retention, performance optimization, and overall readiness in high-stakes operational environments.

ABOUT THE AUTHORS

J.T. Folsom-Kovarik, PhD is a Senior Scientist at Soar Technology, an Accelint company (SoarTech) and researches computer adaptive training systems that assess and adapt to learning needs. Dr. Folsom-Kovarik has contributed in research areas such as user model design, state and trait estimation during user interaction, robust performance modeling, computer understanding of speech and text, activity recognition, and algorithms that enable planning ahead for effective training. Much of this work has focused on adaptation to support individuals and teams in military training contexts. Key outcomes of his adaptive training research include improved user test scores, training time, self-efficacy, and transfer of training. Related research has also advanced cognitive ergonomics, decision presentation and explanation, and user control methods that help to increase the effectiveness, generality, trust, and acceptance of computer adaptive training systems as they interact with humans. Dr. Folsom-Kovarik earned a Ph.D. in computer science at the University of Central Florida.

Angela Woods is a Software Architect at SoarTech with over 24 years of software engineering experience with emphasis on designing system architectures for real-time uses including intelligent training, agent-based simulation, data analytics and game development. She is an expert in dynamic adaptation techniques and her most recent work has included detecting semantically meaningful events in timeseries data streams and using machine learning techniques to build causal models that can augment adaptive intelligent training systems. She excels at coordinating technical solution development on interdisciplinary teams that include both technical and non-technical contributors and has served as architect on over $50M of government funded research.

Daniel Wilson is a Data Science Engineer at SoarTech and develops data-driven solutions to support research in training systems and medical readiness. He is currently pursuing a Master's degree in Data Science at Eastern University and earned a Bachelor's degree in Data Science from the University of Central Florida in 2024. Daniel specializes in machine learning, statistical modeling, data processing, and applied analytics, with a focus on building interpretable and reliable systems that bridge research and real-world applications.

Joseph Cohn, PhD is Vice President of Readiness and Medical Solutions at SoarTech. A retired Navy Medical Service Corps Captain, Dr. Cohn developed and led a diverse range of teams to transition innovative science-driven biomedical and performance-enhancing capabilities across the Joint Force. At SoarTech, Dr. Cohn developed and implement strategic technology development focus for a team of 30 scientists and engineers to deliver Artificial Intelligence (AI) -enabled solutions to enhance human readiness, survivability and health. An expert in developing and executing technology strategies and roadmaps, Dr. Cohn is an Associate Fellow of the Aerospace Medical Association, a Fellow of the Society for Military Psychology and the American Psychological Association.

CDR Lee Sciarini, PhD is US Naval Aerospace Experimental Psychologist # 141 and currently serves as the Deputy Director of Research and Technology Programs at the Naval Air Warfare Center Training Systems Division (NAWCTSD). CDR Sciarini completed his PhD in Modeling and Simulation at the University of Central Florida with a specialization in human systems interaction. CDR Sciarini has conducted and overseen the execution and transition of research in a variety of domains ranging from simulation training system design to neurophysiological assessment of performance.

Beth F. Wheeler Atkinson is a Senior Research Psychologist at NAWCTSD, a Naval Air Warfare Center Aircraft Division (NAWCAD) Fellow, and lead of the BATTLE Laboratory. She has led and transitioned several research and development efforts devoted to investigating capability enhancements for training and operational environments, with research interests in instructional technologies, Human Computer Interaction/user interface design and analysis, and aviation safety training. Ms. Atkinson earned an M.A. in Psychology, Applied Experimental Concentration, from the University of West Florida.

Adaptive Normalization of Assessment Scores: A Multi-Study Validation Approach

Jeremiah T. Folsom-Kovarik, Angela Woods, Daniel Wilson, Joseph Cohn
Soar Technology, an Accelint company
Orlando, FL
Jeremiah@soartech.com, Angela.Woods@soartech.com, Daniel.Wilson@soartech.com, Joseph.Cohn@soartech.com

Lee Sciarini, Beth Atkinson
Naval Air Warfare Center
Training Systems Division
Orlando, FL
Lee.W.Sciarini.mil@us.navy.mil, Beth.F.Atkinson.civ@us.navy.mil

INTRODUCTION

Training and education programs across military and high-performance domains in the world of work rely heavily on evaluation frameworks to ensure learning takes place, to measure performance, and to infer latent traits of learners such as proficiency or aptitude. In this context, *assessment* refers to measuring any aspect of competency (knowledge, skill, attitude, achievement, behavior, and so on) while *evaluation* applies meaning to the assessments in relation to a specific context. For example, the Armed Services Vocational Aptitude Battery (ASVAB) can be used to assess the skills and abilities of test takers (Segall, 2004). The various Services then evaluate the assessment results to determine enlistment eligibility and job opportunities in different Military Occupational Specialties (MOS).

Assessments and evaluation may be criterion-referenced or norm-referenced. *Criterion-referenced* assessments compare performance with conditions and standards, for example to ensure learners reach a required level of knowledge or proficiency at a skill. *Norm-referenced* assessments instead quantify performance relative to peers in a norm group. The norm group are representative individuals who have previously experienced the assessment to establish benchmarks for average and exceptional performance within a certain population. New learners are compared to the norm group to produce a percentile or ranking that relates the individual to the population and enables comparison. Norm-referenced evaluations can inform key decisions such as advancement, career specialization, and leadership assignment.

Norm-referenced evaluation requires that the norm group is representative, the assessment context is similar, and the normed assessments are comparable. The steps to ensure assessments are comparable are referred to as normalization or norming. When assessment differences are suspected, comparison is still possible through statistical methods for removing the differences, called equating the different assessments. However, in long-term teaching and training programs where assessments evolve or undergo structural variation—such as assessment change due to updated pedagogies, operational shifts, or policy changes—traditional normalization techniques may break down and need to be validated to ensure they do not bias the data and decisions. This paper presents a multi-year, multi-study research effort culminating in a robust statistical and AI-driven approach to continuously normalize assessment scores and support norm-referenced evaluation despite structural variability in training programs.

Related Work

In the U. S. military, standardization of teaching, training, and assessment is a key concern to ensure robust and valid selection processes. Researchers use statistical and psychometric analysis to validate instruments used for enlistment decisions (Segall, 2004), selection to officer aviation programs (Williams et al., 1999), and accession to occupational specialties (Parham et al., 2022). A key purpose of this research is to ensure the data being evaluated at the time of selection validly predict long-term success in training and operational job performance.

Norm-referenced evaluation has been widely studied in the field of psychometrics. Most straightforwardly, classical test theory (Traub, 1997) models assessment as observations of a latent trait plus some measurement error. The classical perspective focuses on increasing assessment validity by reducing the error factor. Item response theory (IRT) adds nuance to this understanding, which allows researchers and practitioners to address different sources of

error with statistical tools (Lawley, 1943). For example, various forms of IRT models can account for assessment difficulty, guess rate, sensitivity to differences in a latent trait, and more (Kalkan & Çuhadar, 2020).
The IRT family of approaches provides detail and rigor for norming assessments in an ideal case. However, IRT models are data-hungry and can require hundreds or thousands of datapoints to train a model (Çuhadar, 2022; Svetina Valdivia & Dai, 2024). The selection of sample data is also context-sensitive regarding characteristics of the learners and the assessments (Schroeders & Gnambs, 2024). An example of validating assessments with large amounts of data is the SAT (formerly the Scholastic Assessment Test). New SAT questions are calibrated in part by delivering a few ungraded "pretest" questions (Ali & Chang, 2014) to each of the millions of test-takers—1.97 million last year (College Board, 2024). In contrast, the military setting often does not provide the necessary volume of data for analysis at this low level. Indeed, specialized technical training or leadership development courses may have a few dozen students per class or fewer. We seek an approach that can adapt norming to benefit smaller numbers of learners.

Finally, much of the literature on IRT and statistical norming focuses on written tests, while the military setting often focuses on skill-based performance assessments. This difference has theoretical and practical implications for our analysis. Written tests are easier to design for binary, right or wrong assessments that isolate individual knowledge components. Performance assessment, in contrast, often requires subjective expert review, scores with gradations between good and poor performance, and interactions between skills in context. From a practical perspective, performance assessments in some domains such as ours are also expensive and logistically complicated to deliver. This makes it difficult to apply some standard equating methods such as testing individuals repeatedly to establish test-retest reliability or delivering common items when the curriculum changes to uncover a shared test scale.

We applied and evaluated statistical methods to analyze norm-referenced performance assessment that could model a course with relatively small sample sizes, ordinal graded scores, curriculum changes, and scoring changes over time. This paper contributes a robust, data-driven methodology to analyze similar datasets that can be applicable to numerous military courses.

Motivation and Experimental Setting

In fields such as aviation, medicine, and leadership training, the goal is not merely to determine whether someone meets a minimum standard—it is to identify the best candidates among many qualified individuals. This requires fine-grained and holistic comparisons between students trained under different conditions or at different times. Additionally, high-stakes selection decisions that can affect career progression must be valid, fair, and defensible.

Our experimental domain is a Navy course spanning 28 production weeks and offered at two locations. Learners entering the course have completed specialized screening and prior classroom preparation, ensuring they are already high achievers. The course mission is to prepare graduates for intermediate and advanced training, which entails selection to one of six advanced training pipelines at the end of the course. The Navy selects learners to a specific advanced training pipeline on the basis of norm-referenced criteria, learner preference, and the needs of the Navy.

The main source of performance assessments in this course are approximately 83 training events where learners perform under individual observation to conditions and standards defined in the course curriculum. The exact number of events changes periodically when the course syllabus is revised, and it can also vary when a learner makes more than one attempt or when a training event is skipped or waived.

Each training event represents an opportunity for graded performance of between 12 and 36 required and optional skills whose proficiency level is assessed on anchored, five-point Likert scales. Criteria for progressing in the course establish minimum grades on each skill assessment, and the requirements generally increase throughout the course. Instructors who assess each training event are expert in the operational domain and seek to assess competencies with external validity, relating their grades to long-term training and post-training operational success. They also follow rubrics to increase the interrater reliability or agreement. However, the rubrics are necessarily abstractions of complex skills and do not comprehensively eliminate subjectivity.

Graded skills are weighted and summed to produce a single measure at the end of a course, the Navy Standard Score (NSS). Additional inputs to the NSS, which are not a focus of our analysis, reflect classroom and written tests as well as annotations for any critically unsafe or unprepared training events. The NSS plays a role similar to grade point average (GPA) in college admissions. It allows comparison of learners across time and across factors that might impact

scores, such as the different locations the course is offered. The NSS is a T-score, that is, each learner's raw score is compared to the NSS norm group and scaled to produce a standard normal distribution with mean 50 and standard deviation of 10. The primary NSS is explicitly tied in the course curriculum to selection decisions, in that a minimum NSS is required for certain advanced training pipelines. Relating the NSS to the norm group helps make this selection fair. Learners know that the criterion-referenced performance assessments will be summed and compared to their classmates, which is why they strive to not only meet the grading criteria but consistently exceed them.

With this course as the setting, we reviewed historical performance to author and evaluate a data model. Our data include at least partial data from 4,642 learners recorded between October 2017 and September 2019. The data contain 385,521 training events. There are 3,413 columns (skill grades × training events) that have at least one grade given. In all, this is a challenging number of variables and relationships to evaluate in a selection context. The data also include learner background, primary NSS, and advanced NSS for learners who completed the advanced training.

We classified the learners with these categories:

1. Did not complete both the primary and advanced courses, including those still active in the course
2. Did not complete the advanced course for *non-performance* reasons, such as medical
3. Attritted or recycled from the advanced course due to academic or performance reasons only
4. Completed the advanced course

As a snapshot in time, the dataset contains many learners who are missing assessments outside the selected window. We eliminated learners in the first two categories from the study, leaving N=1,376 learners with primary and advanced course outcomes. Notably, to make the most accurate predictions possible we applied data cleaning to the performance grade inputs, which is a contribution we discuss elsewhere (Folsom-Kovarik, 2025). However, to eliminate advantages our model gets from using cleaned data, we report on all study comparisons using identical inputs.

STUDIES

Over a series of studies, we built and evaluated increasingly sophisticated approaches to address the challenges of comparing student scores, culminating in the repeatable processes and model we present here.

Study 1: Modeling Predictors of Long-term Training Outcomes

Study 1 evaluates the relationship between long-term training outcomes and possible predictors within the early training data. Correlation with long-term outcomes supports the external validity of a measure, as operationalized by the accuracy of model predictions compared to ground truth.

Research Hypothesis
Early training performance can predict long-term training outcomes. (RH1)

Design
Two long-term outcomes are available in the data, completion versus attrition and final NSS after advanced training. We chose to focus on predicting advanced NSS. This is assigned at the end of training, months after learners finish our course. Therefore, it provides a summative measure of training performance where primary NSS is essentially an early indicator. The NSS as a ground-truth measure provides a fine-grained continuum of scores, while completion status is only binary. Advanced NSS is also available for all completers, allowing us to analyze many more students, while only a small minority of learners attrit from this particular program, producing highly imbalanced classes in the data. We use the greater discriminative power of the NSS to compare learners with subtle differences during model development.

We used ordinary least squares (OLS) linear regression to evaluate predictors. To test the predictive power of each variable, we used a separate, bivariate additive nested model for each variable combined with training pipeline assignment as a fixed effect. We selected a nested model because the training pipelines differ in their difficulty. Intuitively, a student with any particular skill level may be expected to earn a lower NSS when compared to the peer group in a more difficult training pipeline.

Results and Discussion

Our analysis provides two statistics to measure external validity (Table 1). First, we report an F-test comparing the model with the predictor against the reduced model with pipeline alone. (The F statistic is followed by the degrees of freedom.) The p-value associated with the F statistic measures whether the candidate early training measures add predictive power compared to the reduced model when predicting advanced NSS. Second, the nested R^2 provides the proportion of variance in the outcome explained by the nested model as a whole. The partial R^2 describes the proportion of variance that each predictor explains in the nested model after accounting for pipeline differences. It can be interpreted as the main effect size of each predictor.

Finally, we also report the magnitude (absolute value) of the maximum regression coefficient. This helps build an intuition for the size of the effect because it is scaled the same as the prediction target, which has mean 50 and standard deviation 10. A coefficient of 5 therefore corresponds to shifting by half a standard deviation in the data.

Table 1: Validity of early training performance as a predictor of the advanced NSS long-term training outcome.

Model	F (DOF)	*p*	Nested R^2	Partial R^2	Max Coef
Pipeline Alone (Not Nested)	7.986 (3)	$p < 0.001$	0.017		2.780
Primary NSS	639 (1)	$p < 0.001$	0.330	0.318	0.813
Number of Waived Events	10.576 (1)	$p = 0.001$	0.025	0.004	1.187
Proficiency Advanced Events	2.512 (1)	$p = 0.113$	0.019	0.002	0.689
Unsatisfactory Events	0.968 (968)	$p = 0.656$	0.013	-0.004	1.248
Number of Performance Reviews	1.005 (968)	$p = 0.481$	0.039	0.022	4.607

We see from Table 1 that the statistical tests reveal a large difference across early training predictors. The first three predictors are statistically very significant, while the last three are not. This tells us that each of training pipeline, primary NSS, and number of waived events correlate with advanced NSS.

Even though the first three rows are significant, the nested R^2 and partial R^2 columns refine our understanding of which is important to model. These show that only one of the predictors, primary NSS, explains much more of the variance in advanced NSS than the others. For this reason, it is important to review both statistics, not the *p*-value alone. Since Study 1 uses a linear regression model, the correlation coefficient *r* is simply the square root of R^2, which produces $r = 0.57$ for primary NSS and $r = 0.16$ for number of waived events. The predictive power of primary NSS is similar to the *r* observed when correlating ASVAB with training completion rates, which is about 0.55 averaged across the Navy enlisted career specializations (Parham et al., 2022). Compared to primary NSS, the small effect size of counting waived events imply that the effect, while significant, is not also substantive.

Finally, the magnitude of the maximum regression coefficient enhances the scale-invariant measure of R^2 with an intuition of the effect size in the prediction target scale. Looking at this column, we see that each performance review a learner requires (due to unsatisfactory performance or progression) can shift the predicted advanced NSS by up to 4.607 points. However, when looking at the table as a whole, we see that the nested model with pipeline and number of performance reviews are not significantly different from the reduced model with pipeline alone, so the larger coefficient should be disregarded.

In conclusion, Study 1 supports RH1 that an early training assessment, the primary NSS, correlates with long-term outcomes and is therefore a valid predictor. We also showed that a linear regression model provides high predictive accuracy (R^2) with a straightforward model structure. We will use this model as a baseline in our final study of an enhanced predictive model. Finally, we discussed how several statistical measures combined can quantify our intuition about which factors are predictive. The F-test comparing models, the R^2 and partial R^2 for describing predictive accuracy, and the regression coefficient combine to highlight one predictor that maximizes all of the measures.

Study 2: Alternative Norm Groups and Assessment Levels

Study 2 first evaluates alternative norm groups. We want to understand the extent to which members of one group have a consistent performance difference in the data. Some group differences, such as differences across locations the course is delivered, can be addressed by creating separate norm groups which make the overall comparisons across groups more fair. Indeed, separate norming per location is already carried out in the existing course.

Second, Study 2 evaluates different levels of parsing or aggregating the performance data. We seek to understand whether levels of aggregation produce significant and substantive differences as compared to the main tool for comparing learners, the NSS. For example, we might wish to compare students on a single skill or a certain timing partway through the course. We must first show that dividing the available data in these ways produces a valid model.

Research Hypothesis
Additional groups divide the data into norm groups that can improve model validity. (RH2)

Design
We follow on the findings of Study 1 by defining our baseline model as a linear regression on both pipeline assignment and primary NSS. We then seek to quantify the extent to which adding grouping factors within the data adds significant and substantive predictive power to the Study 2 baseline. We first examine score differences between cohorts, or groups of individuals who share a common characteristic that is not a performance-based measure.

A second level of analysis addresses primary performance measures that contribute to primary NSS, such as subscales. We investigate the extent to which dividing assessments with subscales adds to predictive power compared with primary NSS alone. We defined eight subscales by grouping skills based on subject-matter expert (SME) knowledge. The assignment of assessments to subscales is a partition (each assessment is in exactly one subscale) and aligns with the course syllabus, enabling SMEs to compare and interpret the subscales. This is an advantage compared to skill grouping with a method such as a principal components analysis (PCA) that would group skills in a statistically optimal manner, but might produce groupings that do not align with how experts understand the course.

As part of determining the best level of analysis for validating the subscales, we also compared different levels of analysis based on timing milestones in the course content, similar to different units of material. Our course is divided into stages, blocks, and training events. The training events that share the same block all target similar graded skills, threshold for progression, and training context. Blocks together make up stages, which each focus on teaching and assessing one larger group of skills. In Study 2, we analyzed the extent to which syllabus units at different time granularity are effective in aggregating the individual performance assessments while producing valid predictions.

If supported by the data, there are theoretical and practical reasons to combine assessments and look for the right level of analysis for combining assessments. First, interpretability is increased (when the factors align with SME understanding, not with a PCA). A smaller number of features improves SME ability to review, understand, and recognize important differences between learners, including during data-driven reviews and selection decisions. Second, grouping assessments provides a smoother data distribution which in our case approximates a normal curve without needing any standardization (Figure 1). The aggregated measures are also more stable when groups have different fields that are missing data. Third, sample sizes in our dataset are only sufficient to support a smaller number of features with a machine-learned model. This is likely to be the case for many military teaching and training domains.

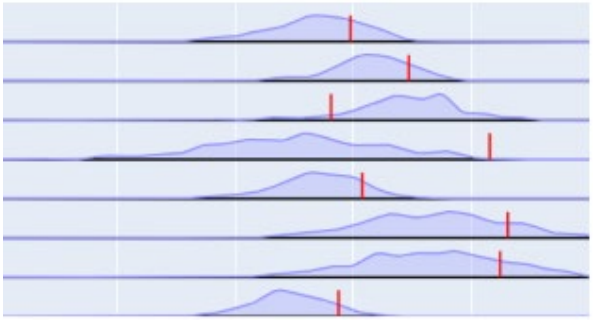

Figure 1: For illustration purposes, eight constructed subscales show how aggregating performance assessments gives an easily interpreted understanding of one individual (red vertical lines) in comparison to the probability density function in a norm group of choice (blue curves).

Results and Discussion
We first examined group differences between cohorts as defined by service branch, syllabus version, and course location. Table 2 shows that from the perspective of the Study 1 evaluation method, the three candidates for splitting the norm group by cohort perform similarly. Service branch and location do provide a significant difference at the $p < .05$ level, while syllabus version approaches significance. However, each of the three groups produces a very small increase in predictive power compared to the model with no split group (partial R^2).

Table 2: Evaluating cohorts indicates that the current norming procedure does correct for group differences as intended.

Model	F (DOF)	*P*	Nested R^2	Partial R^2	Max Coef
No Split Group	639 (1)	$p < 0.001$	0.330		0.813
Service Branch	3.173 (3)	$p = 0.023$	0.334	0.007	2.474
Syllabus Version	3.685 (1)	$p = 0.055$	0.331	0.003	1.095
Course Location	7.806 (1)	$p = 0.005$	0.333	0.006	1.442

According to the Study 1 heuristic, we would use the small partial R^2 to infer that any of the norm group splits make a small difference and are not worth splitting the data at the primary NSS level, considering the threat that reducing the sample size by splitting groups will lead to reducing the robustness and generalizability of the T-score calculation. The small group differences may be interpreted as evidence that the existing norming procedure, which balances for these cohorts, does result in making the primary NSS comparable across groups as intended.

We next evaluate the cohort impact on different levels of assessment data beyond the NSS. When we construct subscales from the raw performance data, which do not already undergo norming similar to the NSS, we see greater difference between cohorts in the divided subscales (Table 3). This increase indicates the need to address group differences when using levels of analysis more fine-grained than the overall, aggregate NSS to predict performance.

Table 3 compares means across cohorts, focusing on course delivery location, which has the largest difference. The table shows how often (as a percentage of valid comparisons) the groups differ in each of the assessments or subscales that contribute to primary NSS. In this table, the Valid column counts the fields (either assessments or subscales) that have mean grades for all cohorts. Invalid fields were discarded. The next column counts the percentage of valid fields with group-based differences at the $p < 0.05$ level. The p_{adj} column counts the percentage of fields whose group difference reaches a corrected alpha value based on 0.05 divided by the number of fields in that row, which in most cases is a much stricter test for significance than $p < 0.05$. The final column reports the percentage of fields where group membership predicts substantive grade differences, producing an R^2 value above 0.05. For each of the three percentage columns in the table, a higher percentage indicates a measure is more affected by group differences.

Table 3: Percentage of assessments and subscales that display group differences.

Assessment Grouping	Timing	Count	Valid	$p < 0.05$	$p_{adj} < 0.05$	$R^2 > 0.05$
Ungrouped Assessments	Training Event	9,476	2,714	39%	13.5%	7.8%
	Block	3,296	905	44%	19.3%	11.4%
	Stage	412	149	67%	45.0%	15.4%
	Course	103	102	67%	46.1%	13.7%
Constructed Subscales	Training Event	736	298	54%	26.5%	9.1%
	Block	256	105	54%	31.4%	10.5%
	Stage	32	15	67%	46.7%	0.0%
	Course	8	8	88%	75.0%	0.0%

The percent with significant differences increases as each timing level aggregates more data – from training event through block, stage, and course analysis levels. Unfortunately, defining levels to aggregate the assessments that are near each other in time (course flow) does not smooth out the cohort differences or make the levels directly comparable across cohorts. However, an encouraging finding is that only the constructed subscales, while following this trend with regard to *significant* difference, show zero percent of fields with *substantive* differences at the $R^2 > 0.05$ level. In other words, aggregating grades using the constructed subscales, unlike timing, does indeed tend to make the score difference attributable to cohort membership small or not substantive. When group differences are not substantive, they may not preclude valid predictions that use more fine-grained levels of analysis than the NSS alone.

The methodology of Study 2 may be used on other teaching and training datasets to understand whether differences between learner groups are significant and substantive at the selected level of analysis and whether they should be addressed by defining split norm groups. Study 2 supported RH2 because we were able to detect groups that should be normed separately at different levels of granularity. We next study methods to correct fine-grained differences.

Study 3: Syllabus Change and Change over Time

We next explore extending the Study 2 method to detect grading differences from syllabus revisions. We expect that changes in the meaning of individual assessment grades can occur when a syllabus revision redefines the context of assessments, when and how they are graded, or adds new assessments. There are hundreds of assessments, each graded on a five-point scale with no fractional values. Compared to Study 2, this requires additional metrics to detect changes and methods to address the differences and enable comparing learners fairly. While we do require an SME to manually complete the tasks of constructing assessments and defining how they relate to course competencies, we can provide automated methods to compare the assessments after they change dramatically or shift over time.

Research Hypothesis

We can detect differences in fine-grained data, such as individual assessments, and adjust the model to account for differences we discover. (RH3)

Design

We evaluated several measures to find which performed best at finding differences in individual assessments. First, the χ^2 (chi squared) test for independence is the most common method of determining if two groups differ in the distribution of grades. However, χ^2 has some limitations. It assumes data are nominal, while our grade data are ordinal. It requires each assessment to have at least five examples of each grade, while in our data some grades are very rarely given. Evaluating more than 100 samples or having more than two outcomes can overestimate the significance of the group differences. We address these points by also calculating (bias corrected) Cramér's V, a measure of correlation between categorical variables that is derived from the χ^2 test statistic (Bergsma, 2013; Cramér, 1946). Filtering for higher values of V ensures that significant group differences are also substantive.

The Mann-Whitney U-test is a nonparametric rank-sum test that can be used to determine if there is a statistically significant difference between the medians of grades (Mann & Whitney, 1947; Wilcoxon, 1945). It requires fewer assumptions than χ^2 and provides more power on ordinal data. This test can indicate the direction and magnitude of the difference using the rank biserial correlation, *r*.

Figure 2 illustrates how these measures work for typical groups with significant differences in grading. The figure shows the distribution of grades on individual assessments. More frequent grades are a wider bulge in the vertical line. (Note that grades are whole numbers, so the curves between whole numbers are only to make the count of each grade visible.) Each pair of lines shows one syllabus or time period in blue and the other in orange. An assessment that is never graded in one group is shown with all grades at zero, and these cases were discarded from Study 2 but retained in Study 3. Cramér's V (left) is high when distributions are very different. Mann-Whitney *r* (right) is high when the second group is graded higher on average than the first group, and negative when the change is in the other direction.

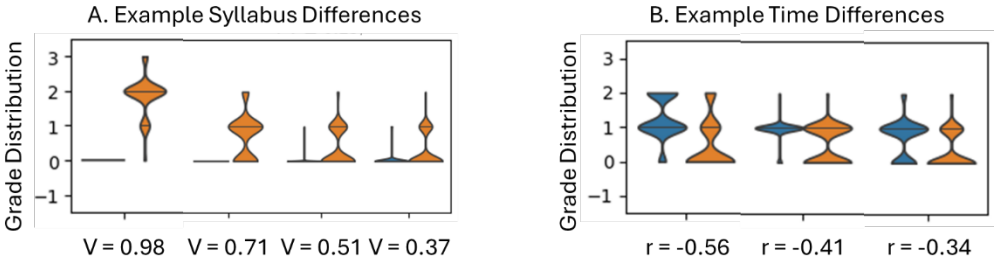

Figure 2: Assessment grading differences visualizing the different values of Cramér's V and rank biserial correlation *r*.

Results and Discussion

We compared these tests to evaluate how many assessments with differences they found. Across assessments that are defined in both syllabi, χ^2 detects 532 out of 839 have significant differences in grading at the $p < 0.01$ level, while the U-test detects 483. Filtering for Cramér's $V > 0.25$ highlighted 122 substantive differences, while 90 significant features had $|r| > 0.25$. Since so many assessments have significant differences, we conclude that it is necessary to find an adjustment for group differences that applies to all assessments. It would add unacceptable error to compare the assessments as if they were not different or to drop all the assessments with significant differences.

We were also able to find 13 substantive changes over time with the same measures, by defining a boundary in time and comparing assessments on either side of the boundary (Figure 2, right). By examination, we found six of the 13 assess the same skill, and eight are within the same stage of the course – a consistent pattern of change over time.

We next explored ways to correct for group differences when comparing assessments. Table 4 shows R^2 (predictive accuracy) of model structures. In the columns of the table, we consider correcting individual assessment grades, both grades and the constructed subscales, and subscales alone. The rows are alternative approaches. The first row shows accuracy without adjusting for grading differences. The second and third rows show that adding groups to the model as a nested factor, similar to the analysis above, actually decreases prediction accuracy. (An additive factor would model the main effect only, while multiplicative includes main effects and interactions.)

In the final row, we evaluate normalizing one group to match another. To do this we collect the mean and standard deviation from one group, and linearly transform the corresponding data from all other groups to have the same mean and standard deviation. We see that this method leads to higher R^2 than the other approaches, but only when applied to the subscales and not to the individual grades.

Table 4: Alternative approaches to correct for group-wide grading differences when comparing learners.

Approach	Individual Grades	Grades + Subscales	Subscales
Unadjusted	.297	.280	.292
Additive Group Factor	.274	.280	.290
Multiplicative Group Factor	.114	.122	.158
Normalize and Scale to One Group	-17.6	-9.78	.315

We conclude that among statistically valid, non-lossy methods of comparing learners across syllabi, the only approach that empirically increases prediction accuracy is to normalize the group distributions. Furthermore, this should be applied only on data that aggregate assessments via a linear method (such as sum or mean, rather than first or lowest value). The subscales are needed in our method because normalizing individual grades, which have only a few discrete values, can produce extreme values in the standard scaler. In particular, the subscales avoid the degenerate case of applying a linear transformation on a zero value when an assessment is missing from one syllabus or the other. Without using subscales these would have to be discarded, losing valuable information.

In conclusion, we demonstrated detecting differences from syllabus revisions and change over time. We corrected for the differences after aggregating by course topic, rather than ignoring them or dropping them. By applying these steps, we were able to correct systematic differences between groups and increase model accuracy. This result supports RH3.

Study 4: Validity and Accuracy of a Predictive Model

Study 4 applies the above results to define an experimental model that predicts NSS at the end of advanced training. We address selection to training pipelines by detecting and handling pipeline differences in the long-term outcome. We apply corrections to address syllabus differences. Finally, we measure the validity and accuracy of the model on the historical dataset, using holdout data and stratified five-fold cross validation.

Research Hypothesis
An experimental predictive model that addresses performance assessment grading differences will increase accuracy in predicting long-term outcomes. (RH4)

Design
We implemented a predictive model with the goal to produce valid, norm-referenced comparisons at the end of primary training and at each course milestone before the end of primary. We compare the experimental model with the baseline from Study 1, the nested model with primary NSS and fixed effect of pipeline assignment, using the R^2 metric. This metric identifies external validity by correlating model predictions with actual outcomes in the historic dataset. Higher prediction accuracy on long-term outcomes increases the evidence that the model correctly predicts learner performance in each of the pipelines, well before the actual pipeline selection decision.

The experimental model involves steps that we implemented based on the results of Studies 1 to 3. Full details may be found in (Folsom-Kovarik, 2025). The experimental model steps are:

- Load and clean the performance assessment data.
- Aggregate performance assessments at the block-subscale level (Study 2).
- Adjust the aggregated inputs for group differences (Study 3).
- Apply a train-test split to the data, stratified by ground truth pipeline assignment (see below).
- Select features based on false discovery rate (see below).
- Measure prediction accuracy and compare to the baseline model (from Study 1).

We evaluated the experimental model and the baseline using a train-test split produced by stratified five-fold cross validation. The pipeline assignment is the stratum, and the folds are selected within each possible pipeline. This is necessary to enable using the model across pipelines for predicting outcomes for each learner in each pipeline, enhancing the primary NSS as a tool for comparing learners before pipeline selection. To support this use case, we show the predictions are valid in each pipeline assignment where we have ground truth. Other groups that also affect model accuracy are not strata, so the model must be robust to their group differences.

The Study 4 methodology aims to detect overfitting the model to available data, which would produce a model that is not robust or predicts inaccurately for new learners. Therefore, R^2 in Study 4 represents the accuracy on predicting only test data points (out-of-sample evaluation). Studies 1 through 3 used all available data in measuring various models' correlation with the prediction target (in-sample evaluation). The train-test split is the reason that the R^2 metrics here are lower than the same metric in the earlier studies.

False discovery rate (Abramovich, 2006) is a machine learning method of selecting input features for inclusion in a model. Selecting a subset of all available features benefits model parsimony and helps avoid overfitting. This method relies on knowledge of the correlations between each input and the prediction target. Inputs that do not correlate with the prediction target are removed, leaving the inputs that together produce a specified, low probability of spurious correlation (which can occur randomly when testing many correlations at once). The features are selected using the model training data and then tested on the remaining data, to mimic real-world application on new learners.

Results and Discussion

First, Table 4 summarizes the experimental model accuracy when compared on the same learner differentiation task the baseline was designed for. The table shows that experimental accuracy improves on the baseline in all metrics. In order, the columns of the table are as follows. Out-of-sample R^2 is as described above, the accuracy in predicting data points that were not in the model training set. In-sample R^2 refers to the model accuracy when trained with all available data and tested on the same data, as was done in Study 1. We see that the experimental model performs as accurately on the test data it has not seen before as the baseline does on predicting the same data used for training the model. The difference between the Pipeline Mean accuracy and Population accuracy metrics reflects two possible use cases. The pipeline mean reflects accuracy across each of the pipelines, which is relevant when we want to select the best pipeline by predicting a different outcome for each one available. The population accuracy is simply measured over the entire historical data, with each learner contributing the same amount.

Table 5: The experimental model predicts the long-term training outcome more accurately than the baseline overall, and the experimental accuracy on test data mirrors the baseline accuracy on the easier training data predictions.

	Pipeline Mean	Population			
Model	**Out-of-sample R^2**	**In-sample R^2**	**Out-of-sample R^2**	**RMSE**	**MAE**
Experimental	0.317	0.363	0.330	9.14	7.13
Baseline	0.279	0.330	0.290	9.35	7.31

Importantly, we next tested the accuracy of the experimental model on only a subset of the course data, which is not possible with the baseline method (Figure 3). We divided the performance assessments in the course according to block and tested the experimental model with only the first block, only the first two blocks, and so on. We found that the experimental approach can produce an increasingly accurate predictive model using a subset of the course data.

There are 32 blocks of performance assessments in our course data. Graphing the change in model accuracy with increasing data, we see that for all pipelines but one (colored lines in Figure 3) the partial data model reaches approximately 60% of the full model's accuracy after five blocks. By around the 20th block, that figure climbs to roughly 90%. This early predictive strength implies that relatively few blocks of assessment data are needed to forecast later advanced outcomes. Finally, all groups reach 100% of the full model's accuracy by the 25–30 block range. This convergence indicates that even though early blocks provide a strong predictive baseline, having the full slate of data ultimately bridges any remaining gaps, ensuring that the partial data model's performance aligns with that of the full model.

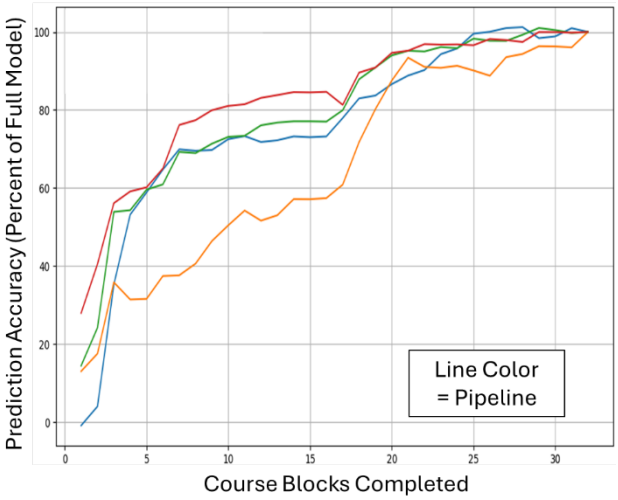

Figure 3: The experimental model produces predictions throughout a course, enabling performance monitoring and early warning. Accuracy increases with more data.

The magnitude and pace of model improvement differ by pipeline. The yellow line indicates one pipeline has lower prediction accuracy between the fourth and 20th blocks of assessments. This trend suggests that later blocks play a relatively larger role for this pipeline, while the other pipelines obtain strong predictive accuracy earlier in the course. Further research is needed before the early mid-course models will be as useful for learners of that pipeline specifically.

We suggest that with proper cautions to users about prediction accuracy, it should be possible to leverage the mid-course model predictions to give learners an insight into whether they are on track with the performance necessary for their pipeline of choice. This is an advantage because the norm-referenced NSS is only available at the end of the primary course. As another comparison, the early model predictions may be easier for learners to interpret than the raw grades that are available in mid-course, because raw grade sheets focus on meeting criterion-referenced thresholds and do not give learners insight into how they compare with an appropriate peer group.

CONCLUSION

This paper shares methods we used to detect and adaptively address possible data issues in performance assessments. We implemented a predictive model that analyzes assessments on fine-grained subscales and timing in the course. The method addresses norm group differences, such as course location, syllabus, and change over time, that would otherwise threaten the validity of the model. The granularity of the inputs at each block of our course allows predicting long-term outcomes during training, using all available data to make predictions in mid-course. Our results support that the model accurately predicts outcomes in a historical dataset. We continue to research and develop the methodology and the experimental model with a focus on increasing predictive accuracy.

Domain Suitability

Based on the analysis, we judge this approach is best suited to environments where normed performance comparison informs high-stakes decisions (Study 1), performances are assessed with structure and metadata that align them to skills (Study 2), and assessment changes incrementally (Study 3). Applicable domains may include medical, aviation training, or technical certification programs.

Excluded domains may include those with fully subjective assessments or inconsistent or missing metadata that relates several assessments to shared, high-level competencies. Manual effort by experts is required to define the competency subscales that unify assessments. Such structure is often available, for example, in the military from training course control documents, occupational standards, or mission essential task lists. In civilian settings, competencies can be drawn from state educational standards, standardized test specifications, or professional certifications. However, if multiple assessments cannot be linked to each competency, for purposes of norming and predictive validity at an aggregate level (Study 3), this approach will not be appropriate.

Implementation Considerations

Technical infrastructure and data management that enable this approach centralized repositories, consistent data schemas, and metadata tagging of assessments with competencies and other course structures. Training personnel to use the system, after the models are created, should involve basic understanding of statistics that describe populations and groups. Training might not require delving into probability, uncertainty, and other topics which may be a barrier to technology acceptance. Finally, while the approach uses statistical techniques to detect and balance sources of bias, a fully implemented system should undergo bias audits, SME validations of inputs and scoring, and subgroup fairness checks to ensure full legal and ethical compliance.

Future Work

The granularity of the model along with the SME-designed assessment subscales suggest that in future work, experimental predictions early in the primary training might be useful for driving early warning about performance shortfalls and recommendations for adaptive training to focus on specific groups of skills for each learner. The methodology might be useful in helping to author syllabus changes by analyzing the likely performance impacts from changing the grading, thresholds to progress, timing, or other context surrounding the course assessments.

Another area for future research is exploring non-grade inputs to predictive modeling. A rich source of data about learners exists in data such as time to complete training events, attempts to reach mastery, hours spent in training, or number of regressions in mastery. It may be possible to relate these to NSS prediction, of course, and also to new inferences about latent traits such as risk-taking or grit that intuitively seem to underlie these additional variables.

Our near-future work will include expanding to new domains to demonstrate the generality of the approach, embedding the model in dashboard tools, and enhancing model interpretability and control using human-AI interfaces. To show the broader applicability of the approach, future work will focus on the competency framework that underlies our model. Competencies link our assessments across syllabus changes and produce the constructed subscales that are the best level of aggregation for comparing individuals. We hypothesize that leveraging general competency frameworks, which describe key knowledge, skills, and behaviors across an organization rather than in a single course, will enable applying the approach to compare individuals from different backgrounds and career paths.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Naval Air Warfare Center Training Systems Division under Contract No. N6134024C0022. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Naval Air Warfare Center Training Systems Division.

REFERENCES

Abramovich, F., Benjamini, Y., Donoho, D. L., & Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, *34*(2), 584-653.

Ali, U. S. & Chang, H. H. (2014). An item-driven adaptive design for calibrating pretest items. *ETS Research Report Series*, *2014*(2), 1-12.

Bergsma, W. (2013). A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society*, *42*(3), 323-328.

College Board (2024). "2024 SAT Suite of Assessments Annual Report: Total Group." New York: College Board.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton Press, NJ.

Çuhadar, İ. (2022). Sample Size Requirements for Parameter Recovery in the 4-Parameter Logistic Model. *Measurement: Interdisciplinary Research and Perspectives*, *20*(2), 57-72.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1-22.

Folsom-Kovarik, J. T. (2025). Predictive Data Analytics to Refine Aircrew Training and Operations. Orlando, FL: Soar Technology Technical Report. January 12, 2025.

Kalkan, Ö. K., & Çuhadar, İ. (2020). An evaluation of 4PL IRT and DINA models for estimating pseudo-guessing and slipping parameters. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(2), 131-146.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, Series A, 23*, 273-287.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*, 50-60.

Parham, S., Held, J., & Blanco, T. (2022). ASVAB Validation Technical Report: Master at Arms (MA) ASVAB Standards for Navy Enlisted Rating Entry. Chantilly, VA: Peraton (Perspecta). May 30, 2022.

Schroeders, U., & Gnambs, T. (2025). Sample-Size Planning in Item-Response Theory: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *8*(1).

Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale*. Seaside, CA: Defense Manpower Data Center.

Svetina Valdivia, D., & Dai, S. (2024). Number of Response Categories and Sample Size Requirements in Polytomous IRT Models. *The Journal of Experimental Education*, *92*(1), 154-185.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement*, *16*, 8-13.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*, 80-8.

Williams, H. P., Albert, A. O., & Blower, D. J. (1999). Selection of officers for US naval aviation training. Research and Technology Organization (RTO) Meeting Proceedings 55 (MP-055). RTO Human Factors and Medicine (HFM) Workshop on Officer Selection. Monterey, CA. November 9-11, 1999.