Item Response Theory Adapts Training to Disparately Skilled Trainees

Courtney Dean[1], Webb Stacy[1], Michael Keeney[1], Eric Day[2], Robert Terry[2], Tom Alicia[3]

[1]Aptima, Inc.	**[2] University of Oklahoma**	**[3]NAWCTSD**
Woburn, MA	**Norman, OK**	**Orlando, FL**
cdean@aptima.com	**eday@ou.edu**	**tom.alicia@navy.mil**

ABSTRACT

Large-scale, distributed live, virtual, and constructive training exercises show much promise in providing effective training, but many challenges must be addressed. By their very nature, these exercises involve individuals who have different experience levels, who perform different jobs or roles, and who necessarily have different training requirements. For any two participants in a common training event, individual proficiency will influence their ability to perform tasks during the event. Interactions or interdependent tasking can result in overwhelming some participants while boring others. As we enhance simulator-based training capabilities with real-time performance measurement and the engineering of adaptable content, there is a clear opportunity to use these capabilities to better serve the trainee. This paper details an approach to leveraging measurement theory and scenario engineering technologies to produce a tool which will inform the "threading together" of scenario experience items based on a diverse set of targeted training objectives.

Specifically, Item Response Theory (IRT), a testing methodology that uses mathematical models to select appropriate test items for individuals based on observed levels of proficiency, can be effectively adapted to simulator based training. Through the selection of appropriate "scenario experiences," or items, IRT can facilitate an adaptive training environment where disparately skilled individuals can train collectively and benefit from training tailored to their skill level. We present trainees with a seamless pre-test, adaptive scenario, and post-test training experience. This procedure develops initial trainee proficiency estimates, provides developmental content based on individual capabilities, and results in an assessment of each trainee's progress. Collective training objectives are met through the presentation of individually tailored content relevant to the targeted knowledge, skills, and abilities. What results is a seamless scenario experience for disparately skilled trainees participating in a team or large scale exercise.

ABOUT THE AUTHORS

Courtney Dean is an Industrial/Organizational Analyst at Aptima. He specializes in the development of performance assessment systems for training effectiveness studies. Mr. Dean received his M.A. in Applied Psychology from the University of West Florida.

Webb Stacy is Vice President of Technology and a fellow at Aptima. He oversees Aptima's current and future technology portfolios. His focus is the intersection of software and computer science with the science, modeling, and measurement of warfighters as individuals and as teams. He has extensive experience in the development of mission critical software, and holds a Ph.D. in Cognitive Science from the State University of New York at Buffalo and a B.A. in Psychology from the University of Michigan.

Dr. Michael J. Keeney has expertise in improving the abilities, motivation, and performance of individuals, groups, and organizations. At Aptima, his work focuses on assessing organizational training needs, developing methods to measure cross-cultural differences, and improving data collection, management, and reporting processes. Dr. Keeney served over 20 years in the U.S. Air Force, with final assignments as an inspector, work center manager, and technical training instructor. He was certified as a U. S. Air Force Master Technical Training Instructor and U. S. Navy Master Training Specialist. Dr. Keeney holds a Ph.D. in Industrial/Organizational Psychology from the University of Akron, a M.A. in Psychology from The University of Akron, a B.A. in Psychology from The University of Maryland, and an A.A. in Technical Management from the Community College of the Air Force.

Dr. Eric Anthony Day is an Associate Professor of Psychology at the University of Oklahoma where he is part of the Doctoral program in Industrial and Organizational Psychology. Dr. Day received his Ph.D. from Texas A&M University in 1998, M.S from the University of Central Florida in 1993, and B.S. from James Madison University in 1991. He has held faculty positions at Valparaiso University and The Ohio State University. Dr. Day's research interests primarily fall in the traditional areas of personnel psychology and human resources management, including job analysis, performance measurement, personnel assessment, and training and development. Much of his research involves training and complex skill acquisition with emphases on cognitive and behavioral modeling, mental models, expert-novice differences, skill decay, and team-based training.

Dr. Robert Terry is an Associate Professor of Psychology at the University of Oklahoma. He has published over 30 peer-reviewed articles and has been a listed investigator on 8 funded grants, including grants from NSF, NIH, and the MacArthur Foundation. He has also served as a member of several federal review panels. Dr. Terry is currently researching measurement and methodological issues in applied psychology. He has written and edited several published articles pertaining to applied statistical analysis, test development, and psychological measurement. Dr. Terry is the author of several SAS software programs for implementing unidimensional and multidimensional IRT analyses. He has also written about the use of IRT methodology in such non-standard measurement settings as sociometry, interrater reliability assessment, and the optimal scaling of tests of emotional intelligence. Dr. Terry received his Ph.D. in Quantitative Psychology from the University of North Carolina at Chapel Hill.

Tom Alicia is a Research Psychologist at the Naval Air Warfare Center Training Systems Division (NAWCTSD) in Orlando, FL. While at NAWCTSD, he has been involved with a wide array of projects as diverse as instructor operator station interface design, performance measurement and after-action review, unmanned aerial systems, medical simulation and training, naturalistic decision making, and physics-based weather modeling. He earned his B.S. in Psychology at the University of Central Florida (UCF), and is currently pursuing his doctoral degree in Applied and Experimental Human Factors Psychology at UCF.

Item Response Theory Facilitates Adaptive Training for Disparately Skilled Trainees

Courtney Dean[1], Webb Stacy[1], Michael Keeney[1], Eric Day[2], Robert Terry[2], Tom Alicia[3]

[1]Aptima, Inc.
Woburn, MA
cdean@aptima.com

[2] University of Oklahoma
Norman, OK
eday@ou.edu

[3]NAWCTSD
Orlando, FL
tom.alicia@navy.mil

TRAINING MUST ADD VALUE

Large-scale distributed live, virtual, and constructive (LVC) training exercises provide training venues in which warfighters can train those skills critical to achieving mission readiness in operational environments. Large-scale synthetic events like the Navy's Fleet Synthetic Training-Joint (FST-J) exercise allow warfighters from many diverse locations to participate from their home stations – saving on travel time as well as airline and billeting costs. Indeed, positive training experiences have been reported by operational trainees and exercise coordinators alike (Jean, 2006; Koon, 2006).

While LVC training events like FST-J show much promise in providing effective training, particularly for joint distributed operations, many challenges still exist in fulfilling that promise. For instance, the difficulty in scheduling such a large-scale, distributed simulation-based exercise is only exceeded by the difficulty in presenting relevant developmental *experiences* during the exercise that provide *effective learning* to the *variety of participants* involved. By their very nature, these exercises involve large numbers of individuals who have different experiences levels, who perform different jobs or roles, and who necessarily have different training requirements. It is not surprising then these varied participants with different needs, in fact, receive disparate training value when exercise scenarios are developed from a one-size-fits-all perspective.

Recent advances in simulation technologies provide trainers with the ability to modify scenarios during exercises. While this is a necessary condition for providing more effective training, it is not sufficient. Unfortunately, these modifications to scenarios are typically made considering the training needs of only a few individual participants, at any given time, resulting in an improved training outcome for some, while actually decreasing the training value for others. For example, suppose in a FST exercise, both a virtual E-2C (a multi-person airborne command and control platform) and an Aegis cruiser (a sea-based command and control and weapons control ship) were providing early warning over-watch, as well as command and control, to the strike group. However, the three-man team in the E-2C found little training value in the event as all C2 tasking was assumed by the personnel on the Aegis to fulfill their training objectives, thereby diluting the value of the exercise for the E-2C personnel.

Adding to this challenge is that of providing training experiences to the participants appropriately based on their disparate skill (i.e., proficiency) levels (e.g., novice, journeyman, expert). For any two participants in the training event, their individual proficiencies in their role-based knowledge and skills will influence their abilities to perform their tasks during the event. In the simplest example, the influence of any differences in their individual proficiencies may be negligible if their roles do not interact. However, in large-scale training events this is rarely the case, and the training events that the participants experience are either overwhelming (for the novice) or boring (for the expert). In another multi-team environment example, imagine F/A-18 trainees being controlled by a Naval Flight Officer (NFO) on a virtual E-2C. In this instance, the NFO is a novice trainee while the two F/A-18 pilots are senior aviators. Because of the proficiency of the NFO, the training events given to this multi-team system must be focused on training the NFO such that the number of contacts he must manage are low and the tracking of those contacts relatively simple. In turn, while the F/A-18 pilots are able to practice their knowledge and skills concerning air-to-air engagement of these contacts, they are relatively underwhelmed by the experiences provided as they could have handled a larger number of, and more complex, engagements.

For these reasons, large-scale, distributed training exercises do not present disparately skilled trainees with as many robust training opportunities as they could. Furthermore, to date, no technology has addressed this training issue, and there has been little basic research performed to determine how best to provide these opportunities in order to provide more effective training to all involved.

INTRODUCING ITEM RESPONSE THEORY

One possible solution is to leverage Item Response Theory (IRT) to adapt training to the needs of individual participants. IRT is a collection of measurement and subsequent applications that statistically model the observed patterns of responses to test items representing underlying latent traits. The IRT process identifies the item equivalent of a scenario experience and presents these experiences at the difficulty level and appropriate order to optimize the achievement of individual, team, and team-of-teams' training objectives across the various proficiency levels of participants involved in the training exercise.

Although IRT was originally developed and has been most extensively used to assess knowledge, underlying latent variables can be any measurable construct. IRT differs from the more familiar classical test theories in that it considers both the characteristics of the respondent and characteristics of the item. Because of this, IRT is independent of the particular sample of both persons and items. This particular strength permits IRT-based comparison across individuals even if they have not taken the same test items, because measurement is based on the underlying construct instead of only the item itself. This benefit to all IRT-based models is critical for a multi-participant environment, as it will permit assessment of individuals across different experiences as well as comparisons of the experiences among themselves.

Although the mathematics underlying IRT can be complex, its basic premises are not. Persons having different levels for a measured trait should respond differently to items (or stimuli) measuring that trait. The degree to which the item measures the underlying trait is expressed as a probability, for a given trait level, to select the correct response. This probability typically predicts that persons low on the trait will be unlikely to select the correct response, and persons high on the trait will be unlikely to choose an incorrect response. Figure 1 presents a typical Item Characteristic Curve (ICC), which graphically depicts how each item performs with respect to person proficiency levels. The y-axis represents the probability of success (P(X|Θ)) on the item. The x-axis represents the latent proficiency level Θ. The level of proficiency is scaled in a z-score metric (such as from -6 to +6). When proficiency levels are low (for example, below -4), the probability of success is near zero. When proficiency levels are high (for example, at +4), the probability of success on this item is near one.

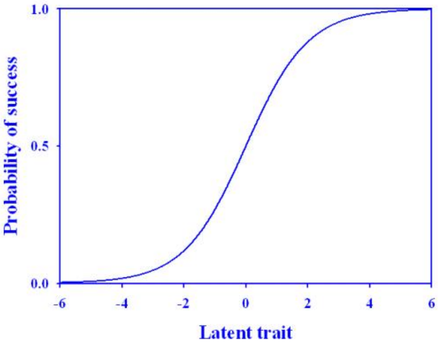

Figure 1. Item Characteristic Curve

IRT models can consider one or more than one dimension. The simplest IRT models, such as one that considers only the one dimension of probability of success, are referred to as unidimensional. Although unidimensional models are conceptually simpler and often sufficiently effective, the reality of assessing trainee performance in a rapidly changing environment requires more complex models. As trainee resources (e.g., working memory) are consumed and as noise (e.g., background conditions) increases, even simple tasks can become difficult to complete successfully. Capturing important features of the environment, independent of the targeted task, is necessary to validly model the task-response probabilities. In this compensatory model a trainee can compensate for a shortcoming in one proficiency by using a strength in another proficiency. Alternatively, a trainee can compensate for a difficult background condition by being particularly skilled in the dimensions needed to solve the task at hand. The choice of future tasks depend not only upon the task difficulty but the background conditions and current status of the trainee – both of which are likely to affect the trainee's future performance.

One challenge in transitioning IRT from a written item format to a scenario environment is to develop a conceptual analog between the conventional test items and exercise inputs. We define an *instructional item* as any change in an exercise stimulus that would require a trainee response. In testing, a test item is typically a written question that provides the stimulus that is presented to the examinee. The test taker responds by choosing a response, typically an option from a multiple-choice or true-false list. Responses are usually scored dichotomously as either correct (1) or incorrect (0), with scores reflecting a total number of items answered correctly. Scores reflect the degree that someone possesses a specified *latent trait* (or *latent factor*) needed to successfully respond to items. For the application of IRT to adaptive exercises the latent traits of interest are the competencies (e.g., knowledge or skills) required to execute the correct actions in response to exercise inputs.

APPLYING IRT TO TRAINING

IRT-driven training consists of two sets of activities, one representing *offline processes* and the second representing *online processes*. Offline processes involve construction of scenarios and instructional items, and the subsequent calibration of those items. Online processes present trainees with the actual adaptive training environment (ATE). Figure 2 displays the steps comprising the offline and online processes.

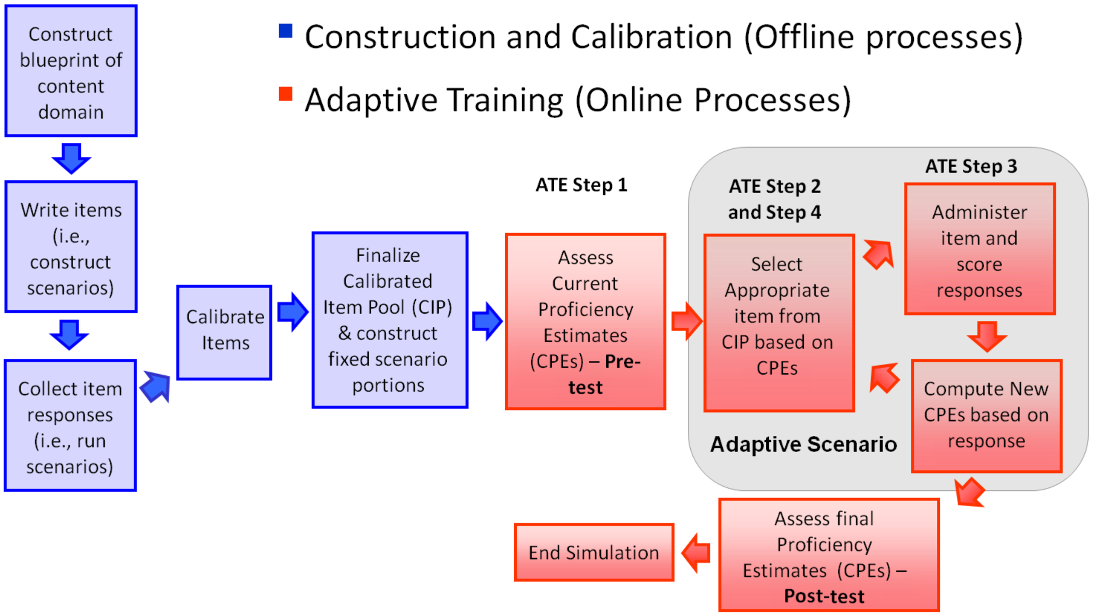
🟦 Construction and Calibration (Offline processes)
🟥 Adaptive Training (Online Processes)
Construct blueprint of content domain
Write items (i.e., construct scenarios)
Collect item responses (i.e., run scenarios)
Calibrate Items
Finalize Calibrated Item Pool (CIP) & construct fixed scenario portions
ATE Step 1
Assess Current Proficiency Estimates (CPEs) – Pre-test
ATE Step 2 and Step 4
Select Appropriate item from CIP based on CPEs
ATE Step 3
Administer item and score responses
Compute New CPEs based on response
Adaptive Scenario
Assess final Proficiency Estimates (CPEs) – Post-test
End Simulation

Figure 2. Overview of Methodology

OFFLINE PROCESSES

The offline process starts with developing a blueprint of the content domain, which provides a template to build scenario experience items. The blueprint guides the development of varied items, ensuring adaptable content to meet the needs of trainees with a variety of skill levels. The blueprint consists of a set of meaningful dimensions that characterize the scenario environment. For illustration purposes, this paper discusses a simple dissection of the NFO role for up to two crew members on an E-2C platform.

Construct blueprint of content domain

Airborne warning and command and control activities can be described by two primary skill dimensions (i.e., competencies): (1) finding, fixing, and tracking (*Find, Fix, Track) of* airborne entities and (2) command and control (*C2*) of assets.

There are a large number of moderating variables that affect the difficulty of successfully executing the critical tasks which the *Find*, *Fix*, *Track* and *C2* skills underlie. It is important to decompose the example items (events) identified in the content analysis of the E2-C performance context in order to identify *a priori* the dimensions most likely to influence item (task) difficulty and then develop items that reasonably sample these dimensions and their possible combinations. The empirical identification of the dimensions that best explain item difficulty translates into empirically identifying the competencies underlying effective task performance. This approach yields strong potential for adapting item content to meet the specific developmental needs of trainees. Not only can this approach better inform the construction of an item library to draw upon for adaptive simulation scenarios, it can also spur the development of an intelligent simulation system that develops novel items extemporaneously.

To guide the development of instructional items (i.e., construct scenario content), we collapse the large number of moderating variables via consensus meetings into a smaller set of meaningful item content dimensions. These dimensions are (1) number of groupings, (2) threat of track groupings, and (3) the ambiguity of grouping profile. Crossing these dimensions in a table matrix, akin to factorial designed experiments, provides a blueprint of item types, which then can be used to guide item (scenario) development. Table 1 shows how these three item dimensions, conceptualized in a 3 × 2 × 2 fashion, yield 12 item types in a hypothetical E-2C simulation. Again, this table can be likened to a blueprint of test content in the development of high stakes tests.

Table 1. Blueprint of Item Types

	One Grouping		Two Groupings	
Threat or not & availability of resources	**Unambiguous**	**Ambiguous**	**Unambiguous**	**Ambiguous**
No threat (neutral)	1	2	3	4
Threat: resources available	5	6	7	8
Threat: resources unavailable	9	10	11	12

Specifically, within a given period of time (item interval), scenarios can be constrained to presenting one or two new groupings or changing the nature of one or two existing groupings. The nature of a grouping can be characterized in terms of threat and profile ambiguity. Threat can be thought of in terms of three levels: (a) no threat (neutral tracks), (b) threat with needed resources readily available, and (c) threat with needed resources readily unavailable (lack of fuel, out of range) or otherwise engaged. Unavailable resources could be made available depending on the officer's actions. Profile ambiguity can be thought of in terms of two levels: (a) unambiguous, which means squawking and flight profile (point of origin, speed, altitude, and vector) objectively indicate a track is neutral or enemy or (b) ambiguous, which means squawking and flight profile do not provide enough information to determine whether a track is neutral or enemy. It should be noted that conditions such as inclement weather can impair radar effectiveness and radio transmissions with consequential increased ambiguity.

Because the E-2C is a team context in which officers must often work together in an interdependent manner to deal with threats and accomplish mission objectives, it is necessary to include interdependency in the articulation of a blueprint of item types. Specifically, interdependency can occur (1) as the vector of a threat from one officer's region of responsibility is shown to be crossing into the other officer's region of responsibility and (2) as one officer faces a threat that the other officer has available resources to handle.

Table 2 shows how adding these two dimensions of interdependency to the previous blueprint of 12 item types yields a 5 × 2 × 2 × 2 blueprint of 40 possible item types. Consequently, Table 2 presents a blueprint

Table 2. Blueprint of Item Types Incorporating Teammate Interdependency

	No Crossing				Crossing			
	One grouping		Two groupings		One grouping		Two groupings	
Threat or not & availability of resources	**Unambig.**	**Ambig.**	**Unambig.**	**Ambig.**	**Unambig.**	**Ambig.**	**Unambig.**	**Ambig.**
No threat (neutral)	1	2	3	4	5	6	7	8
Threat: resources available; *teammate* does *not* have available resources	9	10	11	12	13	14	15	16
Threat: resources available; *teammate has* available resources	17	18	19	20	21	22	23	24
Threat: resources unavailable; teammate *has* available resources	25	26	27	28	29	30	31	32
Threat: resources unavailable; *teammate* does *not* have available resources	33	34	35	36	37	38	39	40

that can guide the development of simulation scenario items that demand differing levels of *Find, Fix, Track* and *C2* skill (i.e., taskwork) as well as differing levels of teamwork (i.e., communication and coordination). The beige cells in Table 2 reflect item types that only demand *Find, Fix, Track* and *C2* skills. The pink and red cells reflect item types that demand *Find, Fix, Track* and *C2* skill in concert with teamwork (the darker the shade of red, the greater the demands for teamwork). In this manner, a blueprint of item content involving differing degrees of needed taskwork and teamwork skills provides a basis for building adaptive team-training simulations that enable the development of successively more complex skills (cf. Kozlowski, Watola, Jensen, Kim, & Botero, 2009).

Write Items

From the blueprint of the content domain, scenario items can be written to likely reflect differing levels of difficulty with respect to the required levels of the underlying skills needed to successfully respond to each item. An item is developed by selecting a cell within the blueprint and generating scenario constraints that reflect the states of the different dimensions aligned to that cell.

Once an item is authored, it can be linked to the competencies (e.g., knowledge, skills) that it is meant to develop. Once linked to one or more competencies, the items can provide information about the trainees' level of skill or knowledge based on their response to the item.

Collect Item Responses and Calibrate Items

The initial calibration phase is so complex and computationally intensive that a real-time calibration of the item parameters is not advised. Therefore, all item parameter calibration is done prior to adaptive training, using calibration samples specifically collected for that purpose. The calibration of experience items is necessary in order to relate the items to a trainee's skill level. That is, until we know how difficult an item is, we cannot use a trainee's response to the item to make any assumptions about the skill level of the trainee. For calibration, raw responses are collected across multiple items to produce an *item-response pattern vector* composed of 0s and 1s. This vector is the raw material to estimate item difficulty (the degree to which successful performance reflects a specific level of the trait of interest), which is then used to choose items for adaptive presentation in a non-scripted scenario.

IRT parameter estimation uses computer-intensive computational methods. The basis for most calculations is to optimize the likelihood function for the data, using either Marginal Maximum-Likelihood (MML) methods or Augmented Bayesian (BAYES) methods (details about these methods are available in Baker, 1992; Lee and Terry, 2006; and Terry and Lee, 2005).

Typical IRT applications, such as high-stakes testing, require precise item calibration (typically a standard error of 0.01 for every item). This high precision requires large sample sizes for calibration, often 1,000 responses for each item. For training purposes, such a high standard is not logistically feasible. It is our position that a standard error of 0.10, which would require a sample of 100 responses, can be both logistically feasible and sufficiently rigorous for simpler IRT models used for adaptive training. We note that larger calibration samples are required as the number of IRT dimensions increases.

Finalize Calibrated Item Pool

Once calibrated items are available, an item matrix can be prepared using the item difficulty calculations. The item matrix, depicted in Figure 3, can be organized by three dimensions: competencies, difficulty and parallel form. Parallel form describes items that target the same competency or set of competencies at the same level of difficulty, but require different environmental constraints. For example, returning to our E-2C example, an item characterized by four MiG 27 aircraft emerging from a strategic red airbase might target the right knowledge and skills at the level for a NFO, but because blue forces already knocked out all of these assets while they were still grounded, their presence would disrupt the reality of the scenario experience. Instead, four SU-32s could be presented (assuming both, that these two combinations of unit type and

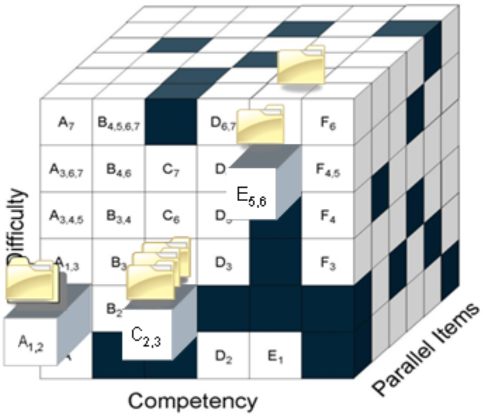

Figure 3. Calibrated Item Matrix

composition are indeed parallel, and these assets are still available). The presence of parallel form items allows for flexibility in the insertion of items into a dynamic scenario environment where progression is often dictated more by student performance than by a master scenario plan.

ONLINE PROCESSES

The online processes involve the actual adaptive training environment (ATE). The ATE can be summarized as a four step process. Step 1 involves the establishment of initial estimates of trainees' skill levels. In Step 2, an item is selected from the item matrix based on the trainees' current proficiency estimates (CPEs). The item is presented to the trainee and performance is scored in Step 3 and Step 4 is a repeat of Step 2. Steps 2, 3, and 4 iterate to adaptively meet the developmental needs of trainees. At the end of the iterative adaptive loop, a final assessment of the trainees' proficiencies is made and the simulation event is terminated.

ATE Step 1: Assess Current Proficiency Estimates

In the absence of preexisting knowledge for trainees' skill levels, establishment of initial trainee CPEs can be achieved through the use of a standardized pretest of pre-selected scenario items. The purpose of the pretest is to assess trainee proficiency levels, thus a testing philosophy is followed. The pretest items are preselected from the matrix at median range of difficulty, much like in adaptive testing. The pretest supports a standardized protocol for assessing trainees' proficiencies at the start of the simulation session. The pretest can be used for formal research and assessment of skill development across a large number of trainees where standardized performance assessments are needed either for research or for student ranking.

The pretest performance of trainees is calculated in real time and a CPE for each applicable dimension is calculated. The preliminary calculation of trainee CPE is based on the difficulty of the items presented during the pretest and trainee's performance on those items.

ATE Steps 2, 3, and 4: The Adaptive Scenario

In Step 3 of the ATE, the adaptive items are administered, responses are collected, and the CPEs of the trainees are updated. As the simulation continues, the ATE returns to Step 3 and the process of selecting adaptive items, administering adaptive items, and scoring responses is repeated to meet the project learning goals or until a pre-determined time period is completed. Thus, Step 4 represents the start of subsequent adaptive iterations in the ATE.

In high stakes computer adaptive testing, items are chosen to maximize the information gain with respect to estimating ability. This is not the primary goal of the ATE. The goal of the ATE is to vary the sequencing and timing of the presentation of items to *optimize learning*. Based on Csikszentmihalyi's (1990; Csikszentmihalyi, Abuhamdeh, & Nakamura, 2005) Flow Theory of skill and challenge and Vygotsky's (1978) Zone of Proximal Development, the strategy of the ATE is to randomly choose upcoming items from a small candidate set of items that are slightly more difficult than the CPE of the trainee. This strategy pushes trainees just beyond their present skill level in order to challenge them without producing performance anxiety. This strategy also lessens the potential for trainees to become bored by the presentation of easy items. Moreover, by choosing upcoming items randomly from a set of appropriate items, item exposure is minimized, thus preventing trainees from "gaming" the simulation after repeated simulation training opportunities.

With two skill dimensions, it is necessary to take note that discrepancies in CPEs between dimensions for an individual should have an effect on item selection. If CPEs are quite discrepant between the two skill dimensions for an individual trainee, the item selection algorithm will weigh the lagging dimension higher when choosing which item to present next. In this manner, the ATE adaptively guides the attention of trainees to their skill development needs without taxing their cognitive resources with explicit feedback and goals midstream in simulation performance (Bell & Kozlowski, 2002; Kanfer & Ackerman, 1989; Kozlowski et al., 2001). Table 3 contrasts two examples showing how the adaptive presentation of items both challenges and focuses attention.

In the case of Team 1, which is composed of teammates with no intra-individual CPE discrepancies, the F-18 pilot would be presented with a more difficult item compared to the item presented to the NFO and the two teammates would be presented with items that have dimension parameter difficulties slightly higher than both their skill dimension CPEs. With Team 1, both trainees would be challenged with respect to both skill dimensions.

Table 3. Adaptive Item Selection and Intra-Individual Skill Discrepancies

		E-2C NFO		F-18 Pilot	
Team		Find, Fix, Track	C2	Target, Engage	Assess
1	CPE	-0.30	-0.30	-0.10	-0.10
	Item difficulty range	0.00 to 0.20	0.00 to 0.20	0.20 to 0.40	0.20 to 0.40
2	CPE	-0.30	-0.50	0.30	0.30
	Item difficulty range	0.00 to 0.20	-0.20 to 0.00	0.00 to 0.30	0.00 to 0.30

Note. CPE and adaptive item are both expressed in terms of a z-scale with values ranging from -3.00 to 3.00.

In the case of Team 2, the NFO's CPE for the *C2* skill dimension is substantially lower than his/her CPE for *Find, Fix, Track* while there is no discrepancy between CPEs for the F-18 Pilot. The F-18 Pilot would be presented with an item that has a dimension parameter difficulty slightly higher than his/her *Target, Engage* CPE but has a dimension parameter difficulty equal to or slightly lower than his/her Assess CPE, while the NFO would be presented with an item that has dimension parameter difficulties slightly higher than both his/her skill dimension CPEs. Thus, with Team 2, the F-18 Pilot would only be challenged with respect to one skill dimension, while the NFO would be challenged with respect to both skill dimensions.

Assess Final Proficiency Estimates – Post-Test

At the end of the training session the CPE for each skill dimension for each trainee is saved for use as a starting value in subsequent training sessions. This approach lends itself to the presentation of a standardized set of items after the ATE as a posttest to conclude each scenario. A final CPE for each participant can be formally assessed through a post-test, as an optional means of determining the change in trainees' skills as a result of the training intervention. Like the pretest, the post-test assessments can be used for research, or for maintaining performance databases for comparisons within or across classes or groups of trainees.

CONCLUSION

Our experience investigating the applicability of IRT to training leads us to believe that it could be leveraged to provide a powerful adaptive training environment for individuals, teams and multi-team exercises. With careful calibration of items and an effective management of the greater scenario context, adaptive items can be inserted into the training environment to provide tailored training to meet each participant's specific needs or to provide the right level of challenge to the multi-team system as a whole. Providing adaptive content to the training environment allows instructors to facilitate improved learning and skill mastery, readying trainees more efficiently and with tangible records of performance.

ACKNOWLEDGEMENTS

The Navy Program Manager for Aviation 205 funded this work. It is managed by the Naval Air Warfare Center Training Systems Division, Orlando, FL. The opinions expressed here are those of the authors and do not necessarily reflect the views of the sponsors or the Department of Defense.

REFERENCES

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, *16*, 87-96.

Bell, B. S., & Kozlowski, S. W. J. (2002). Adaptive guidance: Enhancing self-regulation, knowledge, and performance in technology-based training. Personnel Psychology, 55, 267-306.

Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. New York: Harper & Row.

Csikszentmihalyi, M., Abuhamdeh, S., & Nakamura, J. (2005). Flow. In A. J. Elliot & C. S. Dweck (Eds.), Handbook of competence and motivation (pp. 598-608). New York: Guilford.

Jean, G. (2006). Navy's Virtual Training Exercises Expanding in Realism and Scope. *National Defense*, September 2006.

Kanfer, R. & Ackerman, P.L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. Journal of Applied Psychology, 74: 657-690.

Koon, R. (2006). Navy Links Hornet Simulators on East/West Coasts for Training Exercise. Southern Maryland Online Retrieved March 16, 2008, from http://somd.com/news/headlines/2006/3862.shtml

Kozlowski, S. W. J., Toney, R. J., Mullins, M. E., Weissbein, D. A., Brown, K .G., & Bell, B. S. (2001). Developing adaptability: A theory for the design of integrated-embedded training systems. In E. Salas (Ed.), *Advances in human performance and cognitive engineering research*, Vol. 1 (pp. 59-123). Amsterdam: Elsevier.

Kozlowski, S. W. J., Watola, D. J., Jensen, J. M., Kim, B. H., & Botero, I. C. (2009). Developing adaptive teams: A theory of dynamic team leadership. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 113-155). New York: Routledge/Taylor & Francis Group.

Lee, S., & Terry, R. (2006). *MDIRT- FIT: SAS® Macros for Fitting Multidimensional Item Response.* Proceedings of SAS Users Group International (SUGI), 31, 191-30.

Terry, R., & Lee, S. (2005). *IRT-FIT: SAS® Macros for Fitting Item Response Models.* Proceedings of SAS Users Group International (SUGI), 30, 12-117.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.