finetune

Leading Edge Applications of Advanced Artificial Intelligence and NLP to Assessment

A Finetune Learning White Paper

March 2021

Author: Finetune Learning

Overview

Executive Summary

An essential element of top performing companies and programs using assessments is a well-stocked bank of test content. A healthy collection of test material enables a variety of highly desirable operational and psychometric choices and is instrumental in keeping undesirable risks at a comfortable, manageable distance.

Unfortunately, content authoring is a notoriously difficult, time-consuming, and expensive process. At least two distinct, independent domains of expertise are required in each author: subject matter expertise at the appropriate level and effective assessment authoring skills. The result is that qualified Subject Matter Expert (SME) content authors are often in short supply. Ventures designed to save SME time (e.g., using templates, item engineering, and early Automated Item Generation) have demonstrated that instead of decreasing SME time, additional time can actually be required for developing templates up front, culling the automatically produced items to find acceptable ones, and training SMEs on additional areas of expertise (building item models, programmatic thinking, developing familiarity and facility with AIG tools).

Finetune Learning staff have applied advanced artificial intelligence (AI) and powerful cutting-edge natural language processing (NLP) in a novel way to the problems of test content authoring resulting in a tool called Finetune Generate™ (Generate).

By taking advantage of the significant conceptual domain expertise intrinsic to the product, Generate™ instantaneously creates conceptually and lexically correct test content (including stimuli, stems, and desired item types) designed to preserve the construct representation.

Using this application, SMEs' time is optimized back in line with their primary expertise: iteratively creating novel ways of effectively measuring constructs across domains. The efficiency increase in authoring directly delivers cost containment, better managed test security risk, and revenue to users.

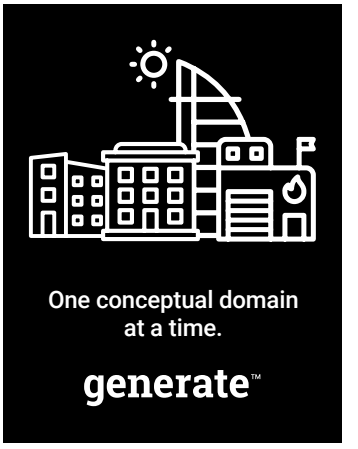
This white paper details the challenges of authoring material and explains how a new paradigm addresses them more efficiently than previous solutions, creating fresh strategic, financial, and operational opportunities for assessment organizations.

One item at a time

Traditional

One item model at a time

Early AIG



Test Development: Challenges and Solutions

Finetune Learning's commitment to education is a foundation of its assessment, learning, and applied advanced technology solutions. From this experience we recognize two important truths. Certain challenges continue to limit the success of well-run organizations that produce assessment instruments; and correctly identifying and understanding a challenge is essential to achieving its solution.

One ubiquitous challenge is that of developing high quality test material. Authoring such test material requires strong subject matter expertise, understanding item/test design and instruction, modeling cognitive complexity, knowing how to maintain construct relevance in a fair and accessible way, and meeting editorial and stylistic requirements. Not surprisingly, training people to have all of these skills is resource and time-intensive, and also demands significant practice to master.

…requires strong subject matter expertise, understanding item/test design and instruction, modeling cognitive complexity, knowing how to maintain construct relevance in a fair and accessible way, and meeting editorial and stylistic requirements.

High quality test material must also have evidence of measuring the construct of interest and doing so within the statistical requirements outlined by the test. Pre-testing or trying out test material involves administering the material to examinees in a format, environment, and under the same conditions that are required during an actual test administration. Testing professionals then use the data to eliminate material that does not meet pass rates (the number of items that pass divided by the total number of items pretested.) The pretesting process is a limiting factor on production, so a substantial pipeline of material to pretest is necessary to meet production requirements.

Having Abundant Quality Material is a Strategic Asset

A limiting factor for the ability to manage assessment production-related problems is how much test material is available. Although the problems differ from each other and come from different sources, having an abundance of high quality test content on hand directly helps solve all of them.

Problems related to Security of Test Material

Item Exposure

As a testing program grows in scale, geographic reach, and construct complexity, the need for more high quality, well-functioning material becomes even more pronounced. If test content is seen too many times or by too many examinees, the material is said to have been “over exposed”, thereby resulting in a greater likelihood of test material being leaked. Incoming examinees may then have the opportunity to memorize questions and correct answers thus defeating the measurement accuracy of the test and the program.

Item Harvesting

The higher the stakes of a decision produced by an assessment the higher the motivation to steal test material and sell it to potential examinees. Despite improved test center security, bad actors still pose as real examinees. Their actual goal is to memorize or take pictures of material so they can sell them — known as “item harvesting”.

Security Breach

In some cases, entire test forms or pools may be stolen and possibly distributed. This is known as a "security breach." If evidence points to this even possibly having occurred, testing programs must err on the side of assuming the material is no longer secure and therefore no longer usable operationally. As a safeguard, companies must over-develop, building additional forms and pools beyond what is required operationally so as to possess "emergency" test forms (lockbox forms) or a separate secure item pool that can be tapped and deployed in case of breach.

Getting Real About the Future Related to Test Security

If we want to be honest with ourselves, we must face the fact that cheating technologies are constantly improving in both quality and in the ability to go undetected. **The fact is that we are already well on the way to the day where there simply will be no such thing as "secure reusable test material."**

Supporting CAT (Computer Adaptive Testing)

The concept of adaptive testing is to administer items 'on the fly' that are neither too easy nor too hard for the examinee based on the correctness of answers as the test proceeds. That is, when an examinee answers an item incorrectly, the test delivery algorithm assumes the material is too difficult and will subsequently administer an easier item. Items answered correctly will prompt delivery of more difficult items. Consequently, adaptive testing requires having enough test material covering content specifications for all levels of proficiency.

In particular, testing programs must take great care to have sufficient numbers of difficult items associated with high scores. If an examinee answers correctly repeatedly (on the way to a high score), each triggers a subsequently more difficult question. If a program only has a small number of these difficulty items, chances are high that those items will be harvested and sold thus putting the entire program at risk.

Providing Practice Material

Practice material is always in high demand. Test material of all kinds and formats featured on the test needs to be available for practice. While retired items and tests are often used, many testing programs want to provide and/or sell additional sets of material to enable multiple rounds of practice. The sale of practice material for test

preparation has become a lucrative business opportunity, making an ample inventory of high-quality items a valuable revenue generating asset.

All of these problems are manageable if ample test material could always be on hand.

Prior Solutions to Meeting Demand

To meet the demand for more material, one solution has been to simply scale up the test development process by increasing the number of people involved. The thought is that more item writers, more reviewers, and more editors could meet the need for increased item production. Unfortunately, efficiency is not typically realized due to overheads associated with more staff requiring substantial managerial effort, time, and expense. This solution also increases operating risk.

In the 1990s, principled item design approaches such as item engineering and Evidence Centered Design (ECD) emerged. These processes, particularly popular in the educational testing industry, ensure that evidence supporting appropriate inferences about specific domains for defined purposes is actually designed into the test material, often by way of relying on well-documented templates and models. The approach enabled high quality material to be produced more consistently and emerge in a way some considered pre-validated for the purposes outlined, while at the same time laying the groundwork for automation.

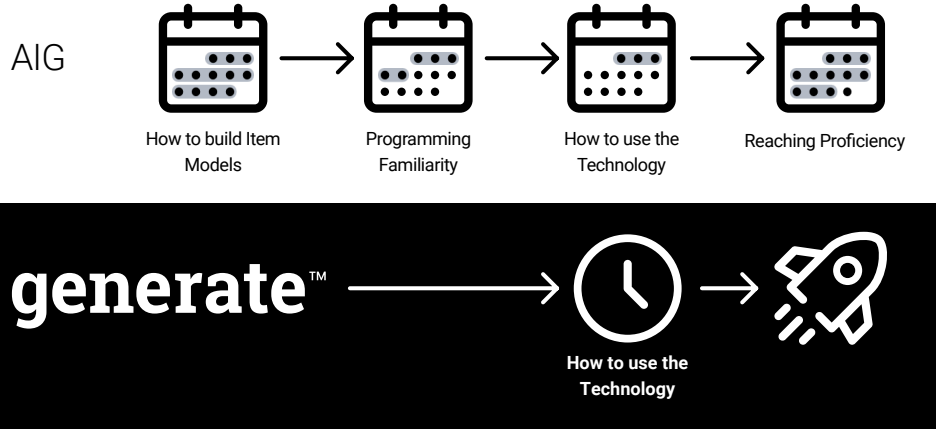
As the use of well-designed templates increased, multiple approaches for how to populate the templates quickly and in more economical ways emerged. Some testing organizations experimented with outsourcing the templates and original item writing first drafts to overseas SMEs to help meet their timelines and presumed cost reduction. Not infrequently, the increased supervisory burden and scarcity of well qualified SMEs called into question the effectiveness of this approach.

The Emergence of Automation — Early AIG

The next approach was to leverage computer automation and the concept of Automated Item Generation (AIG) was born. In this approach, a high-quality item is identified that serves as the basis of an item model or template designed and refined by experts. The expert lays out various dynamic variables that can be instantiated to predetermined variables that have specific characteristics and ranges of values. A

computer program then computes all permutations of alternate values of the variables in the acceptable ranges to create a large number of items with somewhat but limited psychometric variation. Eventually, some artificial intelligence was introduced helping the item models produce items that were even more different from each other. Unfortunately, multiple challenges are also introduced by this approach. A great deal of time, expertise, and many iterations are required to produce the necessary high quality item models. Also, if items coming from a model become known, item variants can be anticipated and practiced, once again meaning that scores coming from the items may not be trusted.

From an SME's perspective, working with AIG tools requires additional skills and training. One must learn how to develop effective item models (some instruction into how programming works has been known to be helpful). Also, significant time is needed to practice building and refining item models as well as time gaining facility with the AIG technology. In addition, a common complaint is that many of the items produced by AIG are not sufficiently close to viable. The result is that so much time is spent sorting and changing generated items that test developers sometimes postulate that the process may have actually been quicker if they had just developed the items themselves. The time, expense, cognitive lift, and practical issues of early AIG tools is experienced on both the front— and back-end of the process.



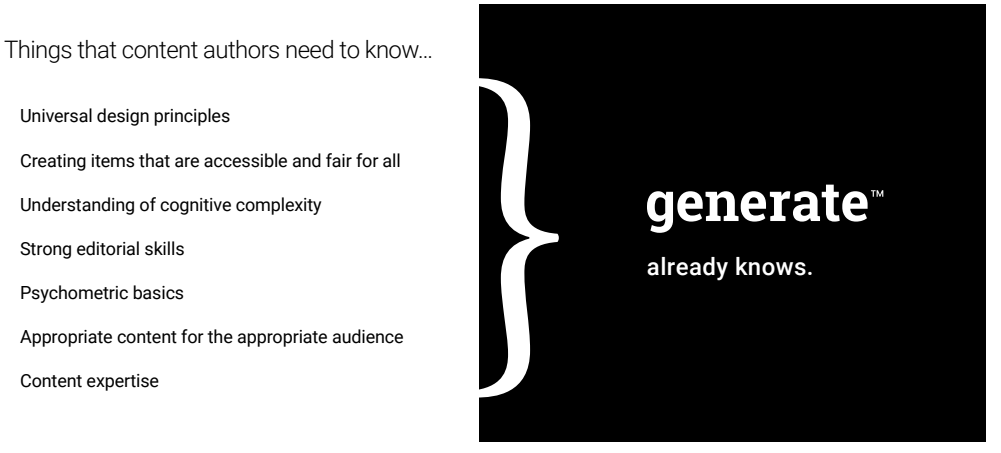
From Automation to Advanced Artificial Intelligence

Experts in learning science, assessment, and psychometrics are collaborating with computer scientists and machine learning experts to make a paradigm leap that goes

well beyond the mere automation of the past. Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence focusing on the computational analysis of natural (human) language data. Advances in AI and NLP are changing the fundamentals of how new test material can be created. Just as so many of us have become comfortable using Google Assistant, Alexa, and Siri to assist in our daily lives, cutting edge AI and NLP is being applied to the specific challenges of personalizing and targeting instruction, learning, and the creation of assessment material.

Finetune Learning staff have applied advanced artificial intelligence (AI) and powerful cutting-edge natural language processing (NLP) in a novel way to the problems of test content authoring resulting in a tool called Finetune Generate™ (Generate).

Natural language processing and natural language generation (NLG) Are at the heart of what allows Generate to learn and understand new testing domains and to generate novel and unique content. Using Language Models and NLG, Generate learns from the lexical complexity of exemplar test material, transforming and building computational representations of language that handle abstract meaning and context. This allows Generate to create new material, mirroring the lexical complexity, semantics, and linguistic construct relevance of the original material.



Generate is built on an architecture that understands the deep structure of test content. It recognizes, analyzes, and represents content specifications including

content maps and outlines, domain semantics, and item formatting. Representative items and their components, such as stem structures, the key, and distractors are understood. The front-end content delivery system integrates elegantly with existing workflows for item review, editing, and approval. Generate is artificially intelligent from the start and with SME verification and feedback it grows and improves with use. Generate understands a wide variety of domains with almost no workflow change while empowering professionals to produce standards–compliant items in standard formats like QTI for seamless integration.

Finetune Generate's advanced AI, NLP and NLG enables it to dynamically create large volumes of well-formed content while substantially reducing item authoring time and effort, sparing SMEs and test managers the extra cost, training, effort, and workflow disruptions associated with the early AIG tools. Finetune Generate is bringing this futuristic capability of "test material on demand" to fruition today.

> Generate enables far more organizations to enjoy the benefits of AI solutions than was ever feasible in the past.

SMEs, Psychometricians, Lamplighters and Typists

Will Advanced Artificial Intelligence obviate the need for qualified test development professionals and SMEs and send them the way of lamplighters and typists? On the contrary.

At the heart of assessment lies the invaluable creativity, insight, energy, and inventiveness of experts: SMEs who write material, test developers who coach SMEs and refine materials, psychometricians who design the measurement model and over-see its effectiveness over time, and managerial leaders responsible for the viability of the program. Generate provides a powerful productivity increase by targeting critical bottlenecks. This allows highly talented, valuable, hard to-find professionals to do more with less, and to do it faster than before.

Current and Future Applications for Generate

Finetune Generate is in use today benefitting high stakes professional licensure and medical certification as well as medium stakes K-12 education assessment organizations. With effectively no disruption to item authoring workflows, Generate has been designed to integrate seamlessly into an organization's authoring process, ensuring that development standards and third-party accreditation important to certification and licensure are fully maintained. For assessment not requiring such a high degree of development rigor, Generate quickly delivers high-quality material with true operational convenience.

High Stakes Professional Certification and Licensure

Certification and Licensure programs typically require fewer items and require high subject domain expertise from SMEs. Legal defensibility is an imperative. The cost of developing items is expensive and key benefits are derived from shortening the most time-consuming phase of item authoring while freeing resources.

High Stakes - Widely Administered Exams

Large testing programs are often the most vulnerable to security breaches. Item drift, intellectual property (IP) security, the need for multiple forms, and lockbox (backup) test forms require large-scale item production targets of high-quality test items, especially if any new items or item types are desired. When new items or coverage is added, there is a need for a large number of new items. This is especially true when adaptive testing is a strategic goal. These programs frequently outsource practice tests to third-party test prep companies. Creating their wholly owned practice tests represents a revenue opportunity that strengthens relationships with test takers.

Old Organizations, New Tests

Many venerable organizations are finding that their traditional tests have become targets of competition from new providers that offer 'good enough / less expensive' alternatives. Improving operational efficiency, reducing production time to market, and rapidly introducing additional revenue generating content (such as improved practice tests) will find great value in improving their competitive position and profitability.

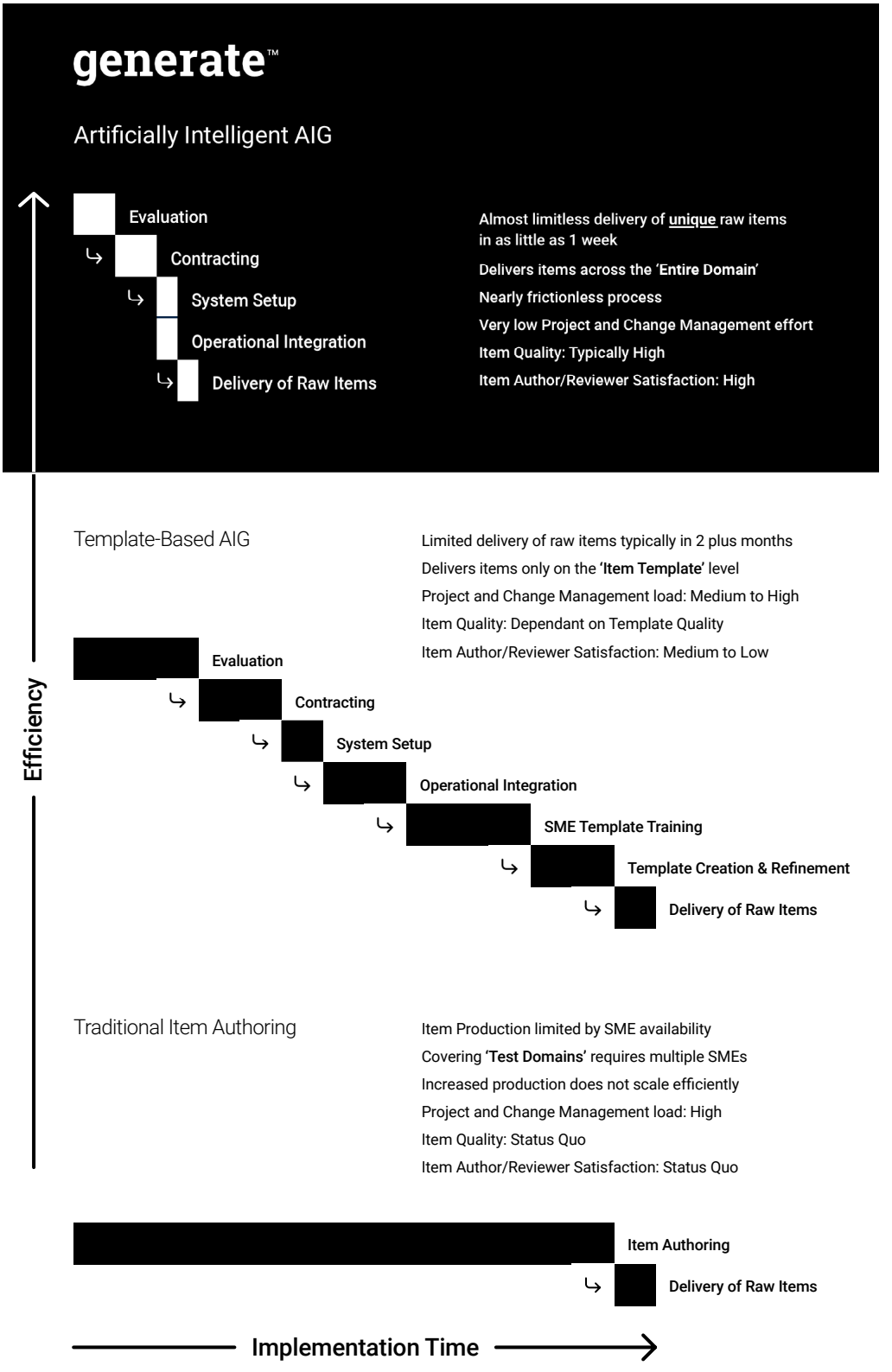
Periodic, Interim, Benchmark and other types of Assessments

Assessments used for formative purposes (and others that are not the actual summative exam) delivered at scale require many of the same characteristics as certifications, including test security and regularly refreshed item content. The capacity to efficiently refresh test content at scale, with speed, and with improved cost-effectiveness represent important functional advantages for programs facing competitive encroachment.

Summary of the New Frontier

Advanced AI insightfully married to state-of-the-art Natural Language Processing provides an exciting, cost-effective, and practical solution to the key challenges routinely faced in authoring that are not well addressed by other approaches. It nearly instantly provides assessment professionals with a wealth of well-formed, intelligent test content at scale. This saves valuable time for SMEs, test developers, and psychometricians while providing cost savings, and business risk reduction.

Every benefit and every opportunity created by a well-provisioned test content bank is available faster and more reliably than was possible in the past. Generate eases staff operational burden while allowing assessment organizations to address common problems. Efficiency and operational resilience are more critical than ever. In addition, Generate keeps getting more intelligent over time without disruptive process changes or more work for staff. Organizations using Generate's differentiating capabilities will be futurizing their test production process.


generate™
Artificially Intelligent AIG
Evaluation
Contracting
System Setup
Operational Integration
Delivery of Raw Items
Almost limitless delivery of unique raw items in as little as 1 week
Delivers items across the 'Entire Domain'
Nearly frictionless process
Very low Project and Change Management effort
Item Quality: Typically High
Item Author/Reviewer Satisfaction: High
Template-Based AIG
Limited delivery of raw items typically in 2 plus months
Delivers items only on the 'Item Template' level
Project and Change Management load: Medium to High
Item Quality: Dependant on Template Quality
Item Author/Reviewer Satisfaction: Medium to Low
Evaluation
Contracting
System Setup
Operational Integration
SME Template Training
Template Creation & Refinement
Delivery of Raw Items
Traditional Item Authoring
Item Production limited by SME availability
Covering 'Test Domains' requires multiple SMEs
Increased production does not scale efficiently
Project and Change Management load: High
Item Quality: Status Quo
Item Author/Reviewer Satisfaction: Status Quo
Item Authoring
Delivery of Raw Items
Efficiency
Implementation Time

About Finetune Learning

Finetune Learning is the market leader in AI-Human hybrid solutions for assessment and education. Trusted by global experts in education and assessment, Finetune creates technology-based products that serve more than 3,500,000 students, 185,000 teachers, and more than 400 schools. Headquartered in Boston, Finetune was a geographically distributed company long before the Covid-19 pandemic, with team members in 14 countries and 10 US states.

For more information and demonstrations of Finetune Generate or Finetune's other leading assessment and education products please visit www.finetunelearning.com or email info@finetunelearning.com for personal attention.

Finetune Generate - instantaneous creation of unique, quality test items
Finetune Catalog - categorizing, indexing, meta-tagging, and aligning content
Finetune Elevate - creation and administration of secure assessments
Finetune Converge - improved inter-rater reliability
Finetune Acumen - consistency in grading student writing

Authors

Saad Khan, Ph.D., Chief Research & Innovation Officer
Finetune Learning

Saad is an experienced tech leader with deep expertise in AI, NLP and Machine Learning with over 12 years of technical leadership developing and deploying machine learning products and solutions to solve R&D and business problems including content generation, recommendation engines, automated scoring, biometrics and multi-modal learning and assessment systems. Winner of 2018 "AI for Social Change" competition and 2020 eAssessment award for Innovations in EdTech he has led global teams of data scientists and software engineers while articulating strategic value propositions to technical leaders, executives and clients.

Sara Vispoel, Ph.D., Psychometrics & Test Development Design
Finetune Learning Leadership Team

For close to 20 years, Sara has led the integration of principled design with psychometric research and test development best practices to develop highly effective, cutting edge assessments across cognitive, SEL, and 21st Century skill domains as well as leading high stakes assessments in K-12, Higher Education and Workforce Readiness. She has managed large teams working on Design; Test Development; Applied Psychometrics; Innovative Item Development; Alignment Studies; PLD development and Standard Setting; Construct Definition and Validation; Eye-Tracking Studies; ACT National Curriculum Survey; ACT Holistic Framework.

Simmy Ziv-el, Chief Business Development Officer
Finetune Learning

Simmy has been a new product strategist and business development leader in the field of education for the past 30 years. From founding his own company through his current position at Finetune, he has worked with globally regarded Research Scientists in turning foundational research into commercially accessible intellectual property through a range of groundbreaking products and services.

Robert Pedigo, MBA
CEO of Pedigo & Associates

Robert has created, managed, and advised companies, government agencies, and international credentialing programs in North America, Asia, and Europe on the creation, management, and improvement of assessment programs, certification management, workforce development, business strategy, and board governance. He has served on nine boards of directors, industry advisory groups, and global standards accreditation organizations.