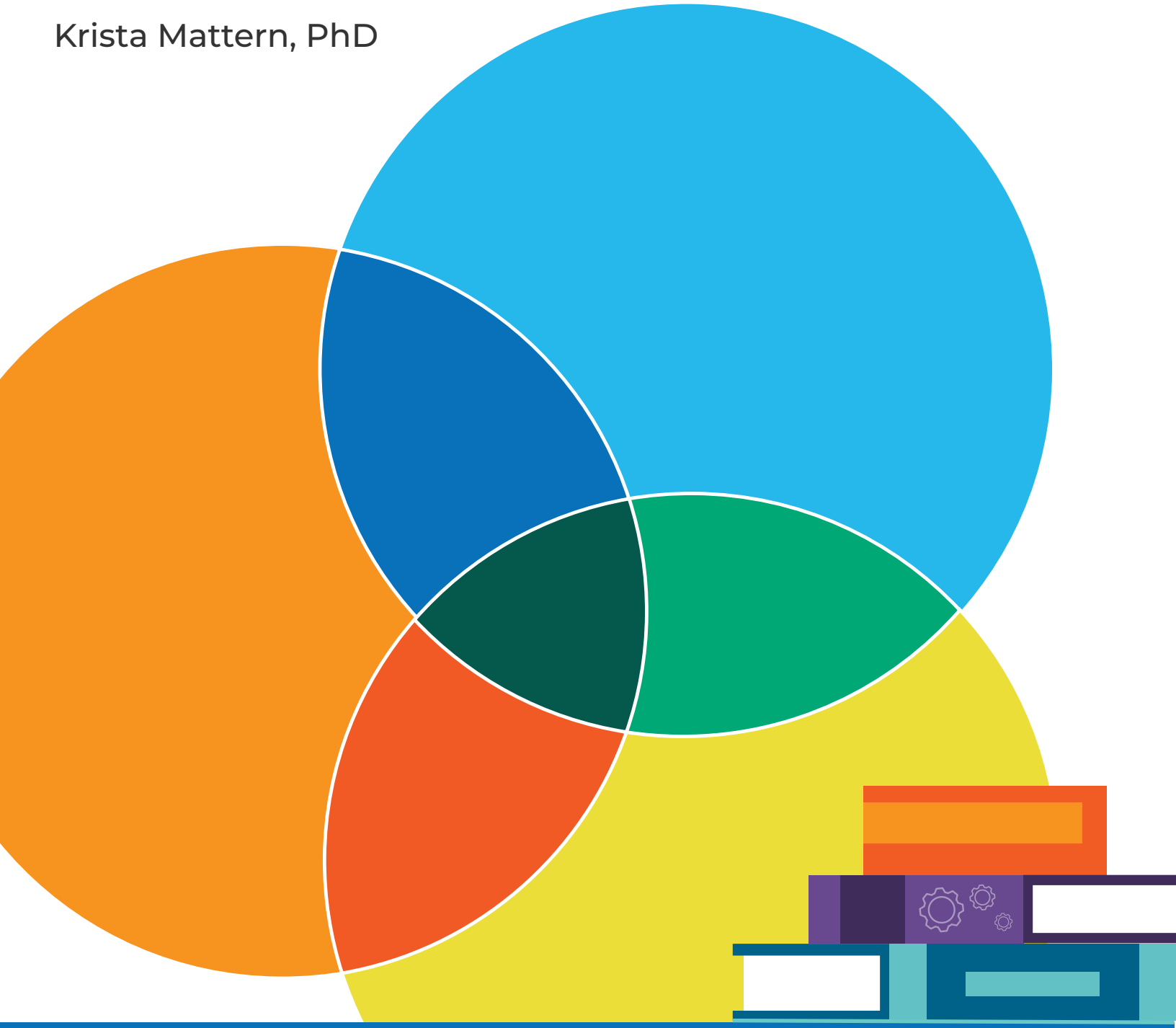
ACT's Efficacy Framework

Combining Learning, Measurement, and Navigation to Improve Learner Outcomes

Krista Mattern, PhD



ACT®

Insights in Education and Work

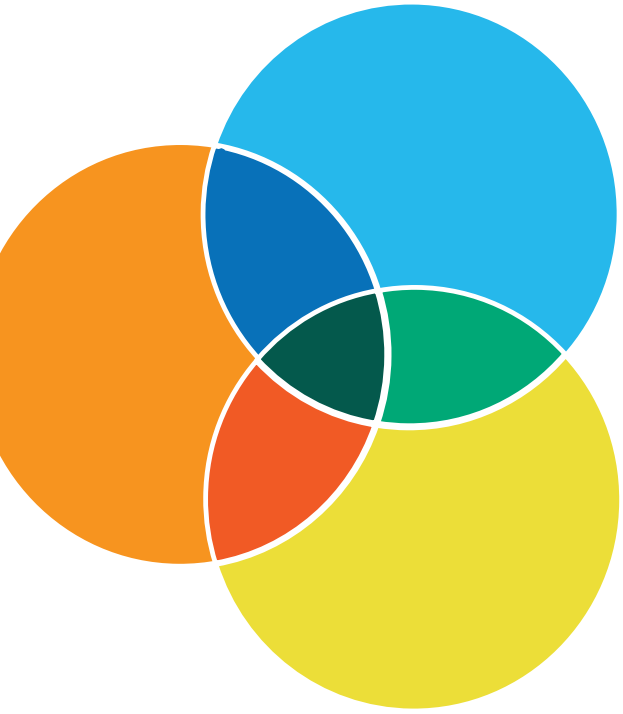
ABOUT THE AUTHOR

Krista Mattern, PhD

Krista Mattern is a senior director in Validity and Efficacy Research whose research focuses on predicting education and workplace success through evaluating the validity and fairness of cognitive and non-cognitive measures. Also known for work in evaluating the efficacy of learning products to help improve intended learner outcomes.

ACKNOWLEDGEMENT

I would like to thank Paul Nichols, Justine Radunzel, Jeff Allen, Sweet San Pedro, Joann Moore, Scott Payne, and Ty Cruce for their insightful feedback and comments on earlier versions of the manuscript.



SUMMARY

In alignment with the *Standards for Educational and Psychological Testing* (2014), which stipulates various sources of evidence can be used to bolster a validity argument for tests, we propose an Efficacy Framework to serve as a blueprint for researchers conducting efficacy research for learning products. In particular, the framework consists of a definition of efficacy and outlines seven sources of efficacy evidence: Evidence Based on User Experience, Evidence Based on Content, Evidence Based on Personalization, Evidence Based on Learning, Evidence Based on Use and Implementation Fidelity, Evidence Based on Relations to Other Variables, and Evidence of Results/Impact. How the seven sources of efficacy evidence are aligned to the four learner outcomes in Kirkpatrick's training evaluation model (1959) is discussed; examples of research activities that may be used to elicit these different sources of evidence are provided. Given that the ultimate goal of a learning product is to improve learner and learning outcomes, we highlight the importance of creating a Theory of Action (TOA) when designing a learning product to increase the likelihood that the intended benefits are realized. In a similar vein, instructional design and learning principles need to be infused in the product design to optimize learning conditions. Finally, an efficacy argument is dependent on the availability and suitability of an assessment to assess current level of knowledge and skills (baseline measurement), which then informs what learning activity is needed (diagnostic/formative feedback), and finally to measure the degree to which the learning outcomes were achieved via the learning product (post measurement). The suitability of that measurement tool is dependent on the validity evidence supporting the use of that assessment to estimate or measure learning. In this paper, we discuss the necessity of integrating both a validity argument and an efficacy argument to create a comprehensive Efficacy Framework.[1]

Equally important to the development of ACT's Efficacy Framework, we have also created a framework for determining the level or rigor of evidence needed for validity and efficacy arguments. This is exceedingly important as an organization must prioritize finite resources as well as consider financial implications. For example, to be competitive in RFPs and other external funding opportunities, it is imperative that research standards are aligned with the criteria of those funding institutions.

The ideas put forth in this document are intended to support and provide thought leadership and direction to efficacy research being conducted, in general and specifically at ACT. At ACT, we are committed to developing and offering products and services that help individuals achieve education and workplace success. This is achieved by researching the appropriateness of their proposed uses as well as their effectiveness in improving intended learner outcomes.

ACT®

Contents

ACT's Efficacy Framework: Combining Learning, Measurement, and Navigation to Improve Learner Outcomes

Krista Mattern, PhD

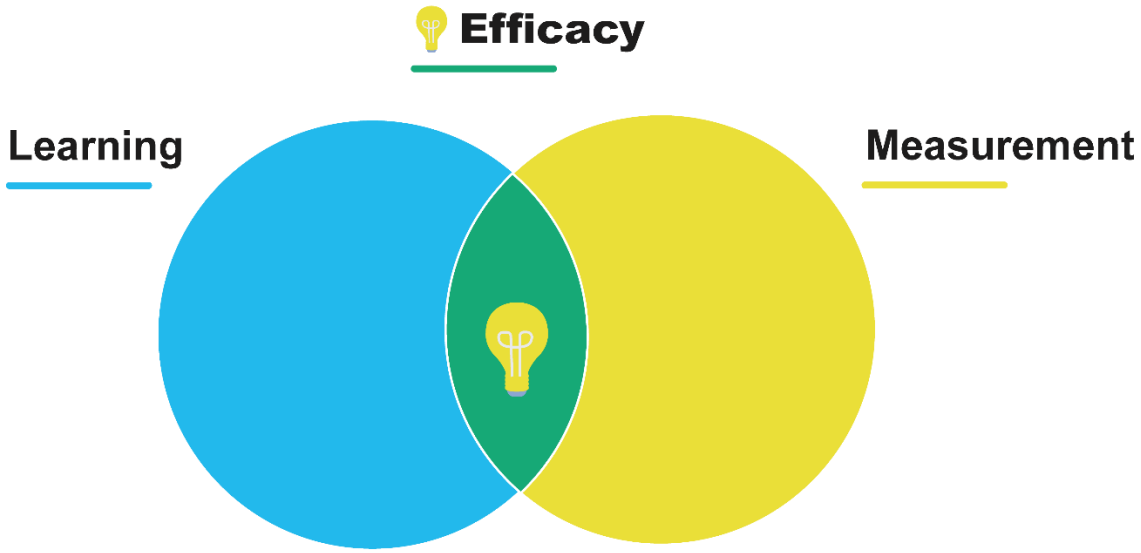
Introduction

A great deal has been written on the topic of test validity. Guiding our work at ACT are *The Standards* (2014), which outlines best practices in test development and validation. As ACT transitions from an assessment company to a learning, measurement, and navigation organization, a framework for our learning products is also needed to guide research activities to ensure that the hallmarks of rigor, science, and scrutiny of our measurement solutions are also built into our learning products. The field of education and education measurement has devoted much less attention to the development of efficacy arguments and frameworks as compared to validity arguments and frameworks. More importantly, the integration of validity theory and efficacy theory into an overarching Efficacy Framework is needed to ensure that we optimize the conditions to stand up an evaluation system with appropriate feedback loops so that we can appropriately assess whether our solutions are designed to have the greatest impact on learner outcomes. This point is exceedingly critical because it lies at the heart of ACT's mission. If we cannot evaluate whether our products are efficacious, then how do we know if we are achieving our mission of helping individuals achieve educational and workplace success? The Efficacy Framework presented here is an attempt to guide research activities at ACT in support of our mission.

In this paper, we advance the proposition that the quantification of efficacy is achieved via the intersection of learning and measurement, as visually displayed in Figure 1. In order to evaluate whether a learning tool is efficacious, we need to measure the impact of the learning tool on the intended learner outcome. This is aligned with the Knowledge-Learning-Instruction (KLI) Framework (Koedinger, Corbett, & Perfetti, 2012). Using a similar visual representation, the KLI framework depicts both instructional events and assessments events as observable factors, which are used to make inferences about unobservable learning events and knowledge components. In particular, instructional events provide or lead to learning events, which in turn leads to changes in knowledge/knowledge components. This change in knowledge is inferred from results of assessment events (e.g., measurement tool).

When selecting a measurement tool to assess learning, Koedinger et al. (2012) identify best practices to support accurate inferences about student learning. First, the measurement tool should be used to assess both short- and long-term retention of student learning to evaluate the robustness or durability of learning (Pellegrino & McCallum, 2017). Second, there should be variety in the task context to ensure that learning will generalize to new situations. Finally, variety in task complexity may be required when integrative knowledge components exist. Similarly, the kinds of knowledge components that a learning product intends to improve should also drive the design of the learning product or instructional features (e.g., repetition, dialogue, explanation, practice) that are most likely to be effective. ACT should use these instructional principles to inform and drive the design of the learning product as well as to inform the development and/or selection of a measurement tool in order to both optimize learning as well as quantify learning.

Figure 1. ACT's Efficacy Framework: Intersection of Learning and Measurement



Along those lines, the ability to collect efficacy evidence of a learning tool is constrained by the availability and appropriateness of a measurement tool (e.g., test) to estimate the amount of learning that has occurred. Therefore, efficacy arguments require both a validity framework for the measurement of learning as well as an efficacy framework for the impact on learning. To this end, this document is organized into the following four sections:

1. Validity Argument for ACT Measurement Solutions
2. Efficacy Argument for ACT Learning Solutions
3. Development of ACT's Efficacy Framework
4. A Framework for Determining the Level of Evidence Needed for Validity and Efficacy Arguments

Validity Argument for ACT Measurement Solutions

As described in *The Standards,* **validity** refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests (2014). In reference to efficacy, we need to collect evidence to evaluate the degree to which test scores support inferences of student learning. For example, the appropriateness of using a specific test to detect whether the learning experience changed the targeted knowledge and skills in the intended manner needs to be evaluated, highlighting the importance of coherence across learning tools and assessments.

Types of Evidence Supporting a Validity Argument

The Standards identify five sources of validity evidence for tests. A description of each of the sources of validity evidence is provided below.

1. *Evidence Based on Test Content* – The degree to which the content of the test reflects the construct(s) it intends to measure.

2. *Evidence Based on Response Processes* – The degree to which assumptions about the cognitive processes engaged by test users occur.

3. *Evidence Based on Internal Structure* – The degree to which the relationships among measurement opportunities and test components conform to the construct on which the proposed score interpretations are based.

4. *Evidence Based on Relations to Other Variables* – The degree to which relationships with other variables are consistent with the construct underlying the proposed score interpretations.

5. *Evidence Based on Consequences of Testing* – The degree to which the expected benefit from the intended use of test scores are realized.

Research Activities and Studies that Facilitate the Collection of Each Type of Validity Evidence

Table 1 lists common types of research studies used to solicit different sources of validity evidence. ACT routinely conducts a variety of research studies in order to construct a validity argument for its measurement products. This section will use the ACT® test for illustrative purposes to discuss different studies that are routinely conducted to provide evidence for the ACT for each of its intended uses.

Table 1. Typical Research Activities Employed to Elicit Different Sources of Validity Evidence

Evidence	Research Activities
Content	Alignment Study
	Curriculum Survey
	Standard Setting
	PLD Workshop
	Job Analyses/Profiling
	Course Catalog Review
	Job Description Review
	Customer Survey
Response Process	Cognitive Lab
	Think Aloud Interview
	Eye Tracking Study
	Item Latency Analysis
Internal Structure	Item Statistics
	Equating Study
	Differential Item Functioning
	Dimensionality
	Measurement Invariance
	Reliability
Relations to Other Variables	Validity Study
	Concordance Study
	Differential Validity
	Differential prediction
	Subgroup Analysis
	Empirical Standard Setting
Consequences of Testing	Efficacy Study
	Effectiveness Study
	Efficacy Trial
	Return on Investment
	Impact Analysis

Content Evidence. The National Curriculum Survey is conducted every three to five years to ensure that the knowledge and skills that are measured on the ACT are aligned with what is taught in high school and needed for success in college (ACT, 2016). This evidence supports the use of ACT for college admissions and placement as well as supports its use for accountability. Additional evidence such as that collected as part of an alignment study, which

quantifies the degree of overlap between what is covered on the ACT and specific state standards, is also a critical component of a validity argument for a state's use of the ACT for accountability.

Response Process Evidence. Prior to operational use, cognitive (cog) labs or think-aloud interviews are routinely conducted to ensure that the cognitive processes that test-takers are using while completing an item are the ones that the test developer. During think-aloud interviews for the ACT, students are asked to verbalize their thoughts as they complete a test item. Such research can help identify items that are unclear, subject to multiple interpretations, or have sensitivity or bias issues. Researchers can use this information to revise or discard problematic items prior to operational use.

Internal Structure. A variety of studies are conducted to ensure that the internal structure of the ACT conforms to the theoretical structure. Analyses to examine reliability, exploratory and confirmatory factor analysis, and differential item functioning are all used to support the internal structure of the ACT. Moreover, given the high-stakes nature of the ACT for college admissions, the consistency of score meaning across forms is of utmost importance. To ensure score comparability, forms are equated so that a score of X on one ACT form means the same as a score of X on another ACT form.

Relation to Other Variables. Given that the primary use of the ACT is for college admissions, a great deal of research has been conducted demonstrating a positive relationship between ACT scores and college success, overall (validity studies) and for relevant subgroups (differential validity and prediction studies). Empirical standard settings are also useful for identifying the ACT score(s) a student needs to achieve to have a reasonable chance of success in first-year college courses. This approach was used to develop the ACT College Readiness Benchmarks (Allen, 2013).

Consequences of Testing. Efficacy studies can be conducted to examine the consequences of testing. For example, an efficacy study could evaluate whether college enrollment rates increase in states that have implemented statewide adoption of the ACT. For a higher education perspective, does using the ACT for college admissions result in an admitted class that is more academically prepared and thus more likely to have succeed on their campus such as earning higher grades and being more likely to persist and graduate. Impact analyses also address another component of consequential validity.

Based on the intended use of an assessment like the ACT, different sources of evidence will be more or less critical to building a compelling validity argument. The evidence that is most critical for the use of the ACT in college admissions is arguably not the most critical evidence needed to construct an argument for using the ACT for accountability. Validity arguments should be framed in terms of the intended use of the assessment. If an assessment is used for multiple purposes or uses, a validity argument for each use should be developed and evaluated according to relevant and existing or newly collected evidence. In particular, if an assessment is used to evaluate student learning, evidence that is most relevant for this use should be collected in support of this use. We will delve deeper into this topic in the following sections.

Efficacy Argument for ACT Learning Solutions

As was presented in the previous section on validity, this section provides a definition of efficacy as well as different sources of efficacy evidence. At ACT, we propose the following definition of efficacy:

Efficacy refers to the degree to which evidence, rationales, and theory support the claim that a learning tool improves intended learner outcomes under ideal conditions.

It is important to note that the literature often makes a distinction between efficacy and effectiveness where efficacy of an intervention is its optimal effect realized under ideal conditions whereas effectiveness is an intervention's typical effects realized under normal conditions. This has implications for our research and reporting the treatment effects of ACT learning tools.

Outcomes Framework

Before evaluating the efficacy of a learner tool, the outcome(s) (e.g., college remediation rate) we intend to impact need to be specified. The outcome of interest can also be framed in terms of a customer problem that we hope the learning product will help solve. For example, higher education institutions may have a high percentage of students that need to take remedial course work and wish to lower that percentage. Kirkpatrick's training evaluation model (1959, 1976) is a popular model in industrial-organizational psychology for evaluating the impact of job training programs on workplace outcomes. Other researchers have highlighted the utility of applying this model to the education space, such as the evaluation of higher education learning outcomes (Praslova, 2010). Given the simplicity and flexibility of the model, we contend that the model can be applied to any learning outcome regardless of the population (students, workers) and setting (school, work). As such, we propose adapting Kirkpatrick's model to research the efficacy of learning products at ACT. As depicted in Figure 2, the model classifies learner outcomes into four categories beginning with more proximal or immediate outcomes located at the base of the pyramid to more distal outcomes located at the apex of the pyramid:

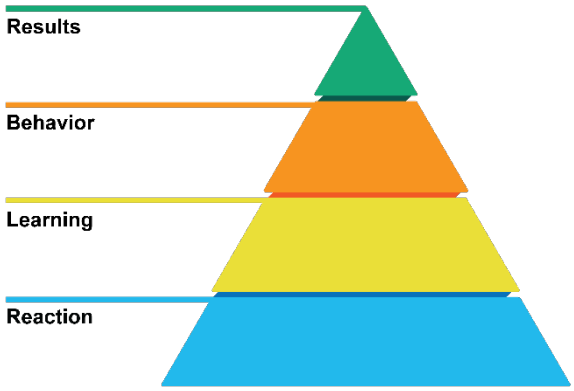
1. *Reaction* – The degree to which individuals find the learning event favorable, engaging, and relevant

2. *Learning* – The degree to which individuals acquire the intended knowledge, skills, abilities, and other characteristics (KSAOs) based on their participation in the learning event

3. *Behavior* – The degree to which individuals apply what they learned from using the learning tool to real-life events (classroom performance, job performance)

4. *Results* – The degree to which intended outcomes occur as a result of using the learning tool. For example, the degree to which increased learning results in learner outcomes we wish to positively influence such as higher college enrollment, lower remediation rates, and persistence and graduation rates.

Figure 2. Kirkpatrick's Evaluation Model



For illustrative purposes, we will use CollegeReady to further elucidate the four categories of outcomes and, in particular, differentiate between behavior and results. CollegeReady is a student success tool that identifies knowledge and skill gaps in math and English and creates a personalized learning path that empowers students to address their individual academic needs. Data shows that many students enter college not prepared for college-level work and therefore are funneled into remedial courses.

With that in mind, the ultimate learner outcome that CollegeReady is intended to improve or the desired "result" would be to increase students' college success, which could be operationalized as lower college remediation rates, higher

college grades, and higher college graduation rates (Stage 4, Results). However, if individuals have a negative reaction to CollegeReady or if they don't find it engaging, then they are less likely to use it (Stage 1, Reaction). If students don't use CollegeReady, then they are not likely to learn the content that is being provided and thus not likely to improve their college readiness in math and English Language Arts (ELA; Stage 2, Learning).[2] If they haven't learned the new content, it is unlikely that students can apply these concepts to new situations, such as performance in a college-level course (Stage 3, Behavior). If we fail to find evidence that CollegeReady is engaging, does not result in learning, and that students are unable to transfer these new skills to new situations, it is unlikely that we will find supporting evidence that CollegeReady is effective at improving the ultimate learner outcome – college success as measured by lower college remediation rates, higher college grades, and higher college graduation rates.

The outcome(s) we wish to impact should be kept front and center when designing and developing new learning tools. In other words, learning solutions should be designed to purpose and not the other way around. Organizations run the risk of creating inefficacious products if new product development is driven by other sources of information or criteria, and validity and efficacy is only an afterthought. Unfortunately, it is often the case that validity and efficacy subject matter experts (SMEs) are not included in the design phase of new product development. Rather, it is common practice that only after a new product is launched that validity and efficacy researchers play a role in the research process, and even then this is a limited role, as they are asked to show that the new assessment is predictive of a certain outcome or the new learning tool is efficacious at improving some learner outcome.

At ACT, we are changing the way we bring new products to market to ensure that they are designed to purpose. In particular, we have implemented a Kanban process for new product development, which clearly defines different stages in the development process and the associated key activities and decision criteria for moving to the next stage. For example, Stage 3 of ACT's Kanban process is Analysis, and it is during this phase that Research contributes to the development of a high-level design (HLD) of the product, including a theory of action (TOA). By stipulating that the development of a HLD and TOA are requirements of the Kanban process, we can better ensure that we are designing new products that are more likely to impact the intended learner outcomes.

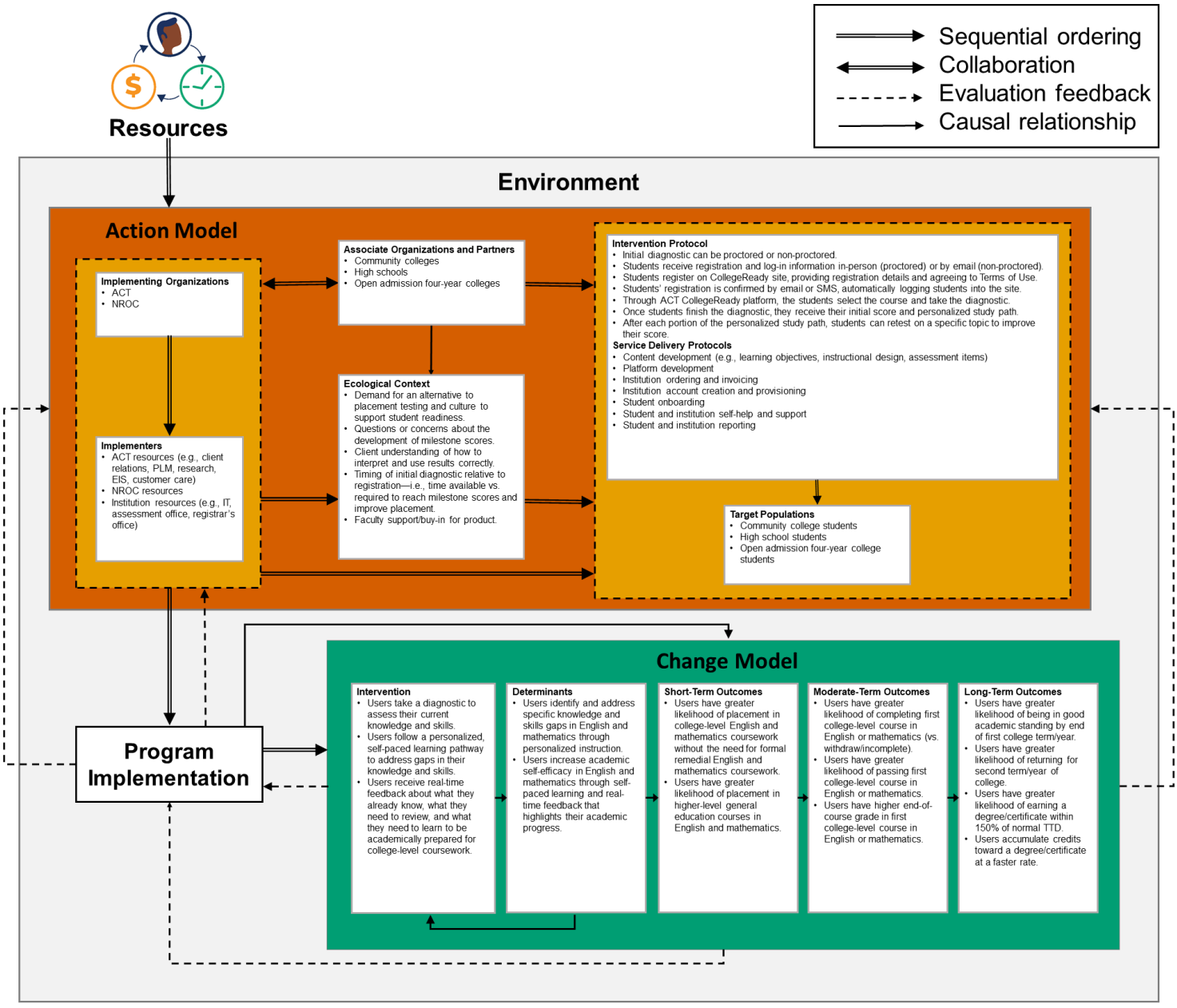
Theory of action

For ACT and its suite of assessments and learning tools, a TOA represents a set of assumptions or a logic model (if-then statements) that describe how an assessment or a learning tool will move individuals from their current state to a desired state (Chen, 2015). In essence, it graphically depicts how the product/learning tool will improve the intended learner outcome(s). When developing a TOA, we start with the learner outcome(s) that we are interested in improving and then work backwards, identifying the intermediate milestones that will get us to the desired end state. Building on TOA models, which focus on the change model or the causal mechanism(s) responsible for the change, we have adopted Chen's (2015) model which incorporates both a change model and an action model. An example of Chen's (2015) model applied to one of ACT's learning products – CollegeReady – is depicted in Figure 3. Given that the effectiveness of different educational interventions are highly dependent on the contextual circumstance under which the intervention is deployed, we believe that combining the action model with the change model provides a more comprehensive and accurate representation of the product ecosystem. The change model helps us understand certain aspects of intervention delivery and its impact on efficacy such as the crucial element(s) of the learning tool, the intended users of the learning tool, and how the

intended users will be contacted and remain engaged with the learning tool.

Chen's Action Model/Change Model also aligns to Kirkpatrick's Evaluation Model. In particular, features of the intervention along with the context in which it is delivered as detailed in the Action Model will impact individuals' reactions to the intervention. The determinant or casual mechanism which has been identified as the cause of the problem informs the content of the intervention or learning product. In this context, the determinant and learning are interchangeable. Finally, the outcomes in Chen's model can include both short-, moderate-, and long-term outcomes, which would encompass both behavior and results in Kirkpatrick's model.

Figure 3. Theory of Action for CollegeReady



Claims documentation

Once a fully detailed TOA is developed, the logical assumptions can be articulated as product efficacy claims. Each arrow connecting two boxes is making an inference about a relationship between two constructs or components of the TOA. For example, in the CollegeReady TOA, a change model is specified depicting a sequential relationship between the intervention, i.e., CollegeReady, and determinants, short-, moderate-, and long-term outcomes. In particular, one of the connections indicates that using CollegeReady will improve their math and English knowledge and skills, resulting in a higher likelihood of being placed directly into college-level English and math courses without the need for formal remedial courses, which in turn should increase an individual's likelihood of passing their courses, persisting to the second year, and completing a college degree.

Each of the claims represented in the TOA needs to be documented, and the degree to which evidence supports each claim needs to be evaluated. If claims are not supported by evidence, the TOA should be revised. If additional claims are being made about the products that are not represented in the TOA, those claims also need to be tested. If supported, the TOA should be revised. If not supported, those claims should be removed from any marketing materials for the product. Desired efficacy claims that are not supported should help inform product improvements and enhancements. ACT's Validity and Efficacy Research department has developed a template to document claims for all ACT products and services. This tool helps identify gaps in evidence in relation to current product claims and thus informs the research roadmap for the product and the priority of specific research studies. This iterative process is represented in Figure 4, illustrating how ACT's validity and efficacy research agenda is set, prioritized, and executed.

Figure 4. ACT's Validity and Efficacy Research Setting, Prioritization, and Execution Process

Types of Evidence Supporting an Efficacy Argument

In this section, we specify seven sources of efficacy evidence. This work builds on the work of *The Standards* (2014) that outlines five sources of validity evidence as described earlier in this document. We adapt those sources of evidence to be applicable to learning products as well as include additional sources of evidence to incorporate Kirkpatrick's training model described above. The seven sources are as follows:

1. *Evidence Based on User Experience* - The degree to which individuals find the learning tool favorable, engaging, and relevant.

2. *Evidence Based on Content* - The degree to which content delivered in the learning tool is of high-quality and aligned to the content of targeted outcome(s), such as course curriculum or targeted standards. Content is in reference to what is covered in the learning resource and measured via an assessment; it is not referencing content knowledge. Therefore, alignment of content is not limited to content/declarative knowledge but could also include procedural knowledge, skills, and abilities and/or practices.

3. *Evidence Based on Personalization* - The degree to which the content delivered via the learning tool is appropriate to the individual's current level of knowledge, skills, abilities, and other characteristics (KSAOs) and adapts as an individual improves their KSAOs within the system. Personalization could also take the form of tailoring content to a student's career and personal interests to promote user engagement.

4. *Evidence Based on Learning* - The degree to which individuals acquire the intended KSAOs based on using the learning tool.

5. *Evidence Based on Use and Implementation Fidelity* – The degree to which the effectiveness of the learning tool is dependent on how it is implemented. What is the most appropriate use of the learning tool? Does the magnitude of learning depend on the implementation model of the learning tool? Does the magnitude of learning depend on other contextual variables?

6. *Evidence Based on Relations to Other Variables* – Degree to which performance in the learning tool is related to targeted outcome(s).

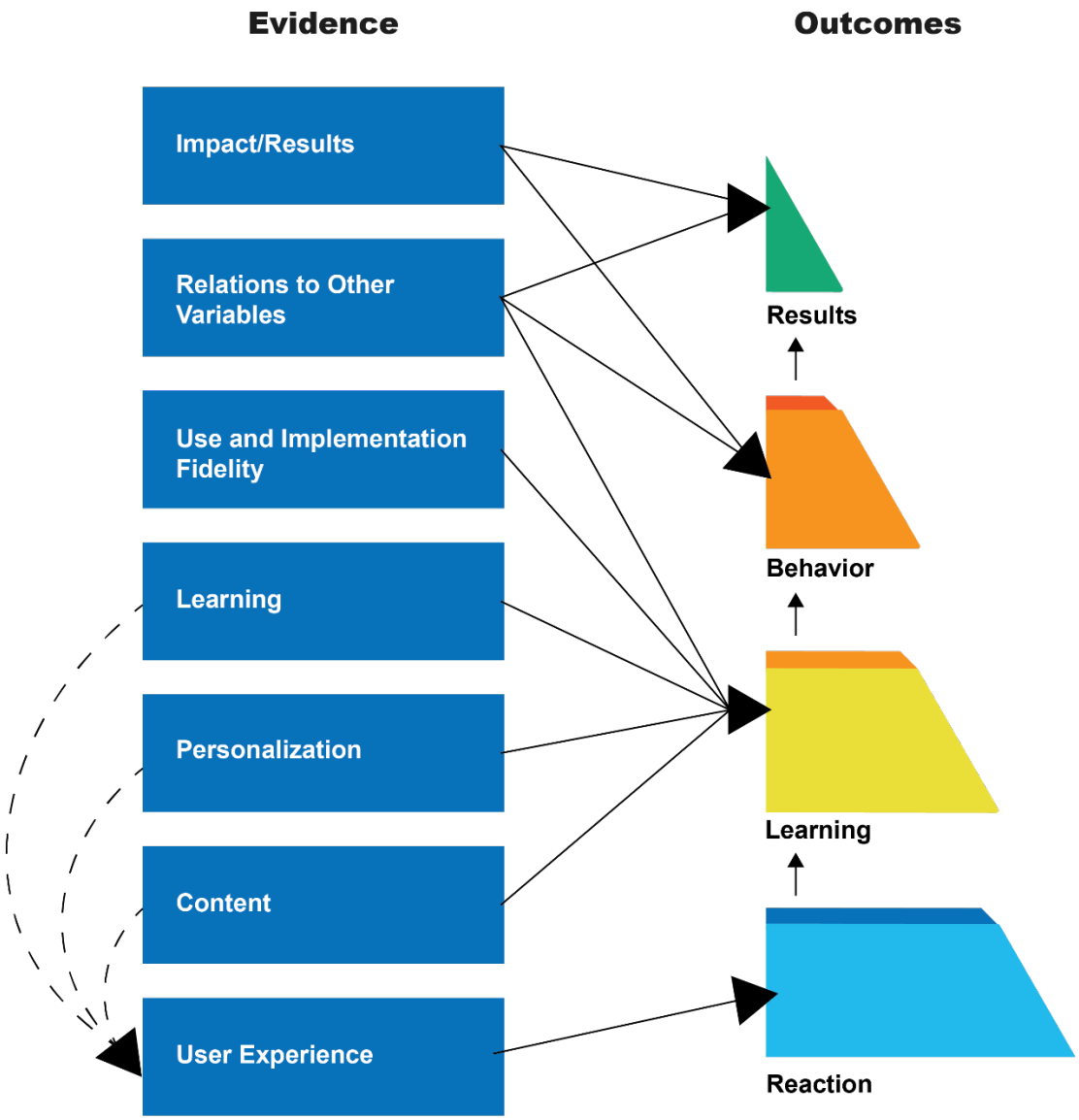
7. *Evidence of Results/Impact* - The degree to which targeted outcomes occur as a result of using the learning tool. Do individuals who use the learning tool have improved outcomes (intended learner outcomes)?

For learning products that make claims about improving learner outcomes that occur far into the future, it is important to test and validate those claims. Given that it will sometimes take months if not years to collect those data, researchers should also plan for short-term efficacy studies where one evaluates the impact of the learning product on intermediate outcomes or potential mediating variables to obtain an early indication of whether or not the product will have a positive impact on these more long-term outcomes. It is important to note that as the time between using the learning product and the collection of results/impact data increases, the likelihood that a small or no effect will be observed increases due to intervening variables, such as subsequent instruction and educational experiences. These additional variables or alternative explanations should be included in the learning product's TOA and incorporated in the design of a longitudinal efficacy study for the purpose of serving as control variables and/or to help contextualize study findings.

Efficacy Evidence and Research Activities Linked to Kirkpatrick's Four Levels of Training Evaluation

For a learning tool, the ultimate goal is that it has an impact on the intended outcome, stage four of Kirkpatrick's model (Figure 2). However, if individuals have a negative reaction to the tool, then they are less likely to use it. If they don't use the product, they are not likely to learn the content that is being provided. If they haven't learned the new content, it is unlikely that they can apply these concepts to new situations (e.g., in the classroom, on a test). This all results in it being unlikely to find supporting evidence that the learning tool is effective at achieving the intended learner outcomes, such as higher test scores or higher college enrollment rates. Given that the degree to which we will achieve success at later stages of Kirkpatrick's model are dependent on being successful in earlier stages, we have mapped the different sources of efficacy evidence to Kirkpatrick's four levels, as shown in Figure 5.

Figure 5. Efficacy Evidence Supporting the Four Levels of Learner Outcomes



As displayed in Figure 5, we are not proposing a 1-to-1 correspondence between the seven sources of efficacy evidence and the four learner outcomes. For example, both user experience evidence as well as content evidence and personalization evidence (indirectly) can provide support that individuals found the learning tool favorable, engaging and relevant. In particular, user experience evidence could include results from focus groups or surveys of existing customers eliciting their satisfaction with the product. If the product is designed to contain relevant, grade-specific or age-appropriate content (content evidence) and that content adapts to a user's current level of knowledge, skills, and abilities (personalization evidence) – that is, it is an adaptive learning tool – the user might find the content more relevant, engaging and less frustrating, which in turn could lead to a better user experience and thus more favorable reactions. Likewise, when a learner experiences positive results by using the product (learning evidence), that is likely to create a positive feedback loop between more positive user reactions, continued use of the product, and more learning.

Figure 5 illustrates how the seven sources of efficacy evidence could potentially provide evidence for the four learner outcomes in Kirkpatrick's model. As was provided in the validity section, we have also provided example research activities that may be used to elicit these different sources of evidence (refer to Table 2). This section can help inform the types of research activities or studies that efficacy researcher may want to consider in terms of collecting specific types of efficacy evidence in order to bolster an efficacy argument for its learning products. This section will focus on ACT Online Prep (AOP), ACT's test preparation product to help students prepare for the ACT, for illustrative purposes to discuss different studies that may be conducted to provide evidence for the efficacy of AOP.

Table 2. Research Activities Employed to Elicit Different Sources of Efficacy Evidence

Evidence	Research Activities
User Experience	A/B Testing Cognitive Lab Think Aloud Interview Eye Tracking Study Focus Group Customer Survey UX Research Game Design/Gamification Rapid Prototype Testing Net Promoter Score
Content	Alignment Study Curriculum Survey Standard Setting Job Analysis/Profiling Course Catalog Review Job Description Review Customer Survey
Personalization	Model and Algorithm Design Algorithm Development & Iteration
Learning	Learning Principles/Instructional Design Efficacy Study Efficacy Trial
Use and Implementation Fidelity	Implementation Study Effectiveness Study Case Study
Relations to Other Variables	Predictive Validity Study Differential Validity Differential Prediction
Impact	Efficacy Study Efficacy Trial

Evidence Based on User Experience. As mentioned above, user experience evidence for AOP could be collected by surveying existing customers. The survey could collect information on what customers liked and disliked about the product, the extent to which they thought the product helped them prepare for the ACT, and whether or not they would be likely to purchase the product again or recommend it to a friend

(net promoter score). In cases where a product has not yet gone to market, usability testing, A/B testing,[3] or rapid prototyping methods could be employed to incorporate targeted customers' feedback into the product prior to launch.

Praslova (2010) indicates that reaction data is the most common data collected when evaluating a training or learning intervention because this is the easiest data to collect. She stresses the importance of collecting other sources of training outcomes data. We agree that the other sources of data are crucial but we do not want to undermine the importance of collecting reaction data. Given the recent explosion in the availability of online learning tools, the market for learning products can be classified as a "competitive" market. That is, if a user is not satisfied or engaged with our learning product, they are going to move on; there are plenty of other options out there. We no longer have the luxury of a captive audience as we do with standardized tests such as ACT but rather we have to captivate our audience. Secondly, and more importantly, positive reactions to a learning tool is a prerequisite for engagement and usage and hence of learning. For instance, a lack of efficacy evidence for AOP may indicate that the product has poor learning content and instructional design or it may indicate that the product's user interface is difficult to navigate. Knowing not only if but why the product is not achieving the desired outcome is necessary evidence to drive product improvements and enhancements.

Evidence Based on Content. For evidence based on content, an alignment study could be conducted to ensure that the content provided within AOP is aligned to the content tested on the ACT. For example, we would want to be able to show that the math knowledge and skills that are taught in AOP align to the math knowledge and skills tested on the ACT. Stronger alignment between the content delivered in the learning tool and the intended learner outcome(s) should translate to more learning.

It is important to make a distinction between learning versus performance gains and the role of content alignment. It is possible to see large performance gains (e.g., large score increases) without learning. Students may have high retrieval strength for particular facts due to recent exposure to the material but low storage strength (Bjork & Bjork, 2011) where students have failed to achieve robust learning (Koedinger et al., 2012). That is, a student may do well on an exam if he crams the night before but may not be able to retrieve any of the newly "learned" content a week later. We don't want content alignment to be confused with a desire to have high retrieval strength or to promote rote learning. Rather, the point is that content alignment is important for efficacy research in that the knowledge and skills that we want students to learn should be appropriately aligned between the learning product and the assessment of learning to be able to detect whether the learning product was effective. If a learning resource was designed to improve students' mastery of algebra, their performance on a geometry assessment may not be a good indicator of the effectiveness of the learning product – hence, the need for strong content alignment. On the other hand, the features and implementation method of the learning product should be designed to promote true learning (storage strength/robust learning) versus short-term performance gains (retrieval strength) by incorporating best practices based on learning sciences research, such as varying the conditions of practice, spacing study or practice sessions, interleaving versus blocking instruction, and using tests as learning events (Bjork & Bjork, 2011) as they align to the type of knowledge change desired as well as the type of learning and assessment events being employed (Koedinger et al., 2012).

Evidence Based on Personalization. Whereas ensuring that the content of the learning tool is aligned to the learner outcomes can promote learning, evidence of personalization can amplify that learning. One component of AOP is a personalized learning path. The degree to

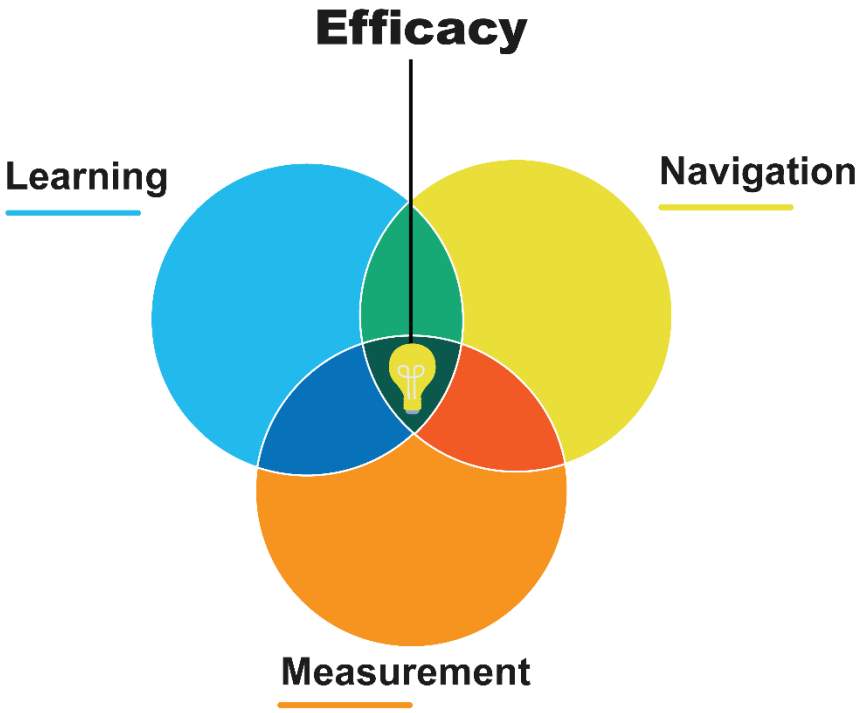
which the algorithm that was developed to determine personalized content for users is working appropriately in the sense that it accurately calibrates students' current level of knowledge, skills, abilities and other characteristics (KSAOs) and adapts with the user as they learn will determine the extent to which learning will be amplified. A study that shows that students who follow the personalized learning path experience larger ACT scores gains as compared to students who use the self-directed option, holding all else constant, would constitute evidence supporting personalization. Other research activities may include algorithm iteration and refinement to maximize adaptive learning.

Personalization could also be in the form of serving up content that aligns with one's career and personal interests to foster engagement in the learning process. It's quite likely that the ways in which we can personalize learning products in the future will take on additional dimensions or forms as more behavioral data becomes available and we can use that information to best tailor learning events to individuals' needs. For example, in the future we may want to tailor communications (e.g., nudges) between the student and the learning product that capitalize on the student's personality profile or social and emotional learning skills to maximize product usage and thus learning. We may also want to provide personalized feedback and insights that take into account not only how the student performed within the learning product in general but also how they performed in relation to their educational goals and plans.

By introducing the concept of adaptive learning, we propose that the overlap of ACT's three strategic pillars – learning, measurement, and navigation – not only constitutes efficacy but represents the amplification of efficacy, as displayed in Figure 6. We have already made a case that the intersection of learning and measurement constitutes efficacy (refer back to Figure 1). Moreover, adaptive learning – an area that is gaining traction in the online learning space –combines learning and measurement using technology and platforms that bring together the act of learning and teaching, the measurement of that activity, and the criteria framework that defines it. Such a system enables, accelerates, and validates learning. Learning analytics, which involves the collection and analysis of a learner's holistic data for the purpose of understanding individual or group interactions and activities and optimizing learning experience by leveraging big data, analytics, and empirical research and insights on all aspects of education, also plays a role in this area.

We argue that both adaptive learning and learning analytics support ACT's strategic pillar of navigation, which encompasses the ways that ACT can connect with customers to provide guidance and navigation on their journey through life. Solutions help individuals and their advocates (parents, counselors, employers, etc.) make good, informed decisions. According to this definition of navigation, customers are assumed to be an active participant in the navigation process at more of a macro level. For example, ACT data can help individuals choose appropriate learning products or apply to a good-fitting college or declare a major that is aligned with their interests. However, we can also conceptualize navigation at a more micro and passive level. Through the development of adaptive learning algorithms, we guide or navigate students – perhaps unbeknownst to them – through a sea of learning content and deliver the resources that our research has determined are the most relevant for them and continues to adapt based on their learning, serving up the next most relevant piece of content. Therefore, within the closed system of a learning platform, we ensure students are maximizing their time and experience and thus hopefully their learning by focusing on content they are prepared to master and filtering out content they have already mastered or not yet prepared to master (e.g., too challenging).

Figure 6. Amplifying ACT's Efficacy Framework: Intersection of Learning, Measurement, and Navigation



Evidence Based on Learning. An efficacy study would need to be conducted to evaluate the degree to which the use of the learning tool such as AOP resulted in learning. For example, a research question might be whether students who use AOP improve their math knowledge. Efficacy studies typically employ a pre-posttest study design. Supporting evidence would show that students who used the product showed more growth from the pretest to the posttest than a control group. Performance gains on the ACT could be used as evidence as student learning. At this point, readers may be thinking that ACT performance gains represent the ultimate outcome, or what is referred to as impact (level 4), and not learning. We contend that improved ACT scores is a means to an end and not an end in itself. We will include ACT score gains in the description of impact evidence but will also discuss more distal learner outcomes that we hope to impact (and are articulated in AOP's TOA) such as college enrollment and college success.

Evidence Based on Use and Implementation Fidelity. As referenced in the distinctions being made between efficacy and effectiveness and the incorporation of an action model (in addition to a change model) in our development of TOAs, evidence based on use and implementation fidelity is exceedingly important as learning products become more and more unstandardized through online delivery and individual discretion in terms of how, when, and where the learning tool is used. Analysis of AOP usage data could provide insights into the extent the tool is being used in practice. For example, our research shows that a large percentage of AOP users utilize the product quite infrequently. Follow-up analyses could examine whether there is a relationship between duration and/or profiles of AOP use and ACT score gains (e.g., using the product for more days results in larger ACT score gains), which could inform recommendations around best practices in terms of optimal usage. The low usage could also provide evidence for why larger effects were not found and should prompt

a deeper investigation of the product usage data and why usage is low.

For our business to business customers, implementation studies could be conducted where we survey different districts or schools who purchase AOP for their students and capture information on how they are using it. For example, are some schools devoting class time for students to use AOP? If so, how are they using that class time? Is it simply like a study hall or is the teacher facilitating and providing some sort of instruction? If instruction is provided, what is the approach used? Are other schools instructing students to use it on their own time? We would want to examine whether different implementation strategies are more or less effective at improving the intended learner outcomes. Additionally, this line of research would also want to test whether different components (i.e., flash cards versus videos) are more or less effective at achieving the intended learner outcomes. Such findings could help highlight best practices and provide guidance to future customers on the most effective ways to implement AOP at their school or district as well as drive product improvement.

Evidence Based on Relations to Other Variables. Validity studies documenting a correlation between AOP performance and ACT performance is a critical source of evidence to demonstrate that we can appropriately evaluate the degree to which learning has occurred. For example, do students who correctly answer practice questions at a higher rate in AOP tend to have higher ACT Composite scores? This source of evidence is a pre-requisite to being able to demonstrate that learning has occurred. If AOP performance is unrelated to ACT performance, then improvement in AOP performance through the use of the learning tool would not translate to improvements in ACT performance. As stated at the beginning of the report, the ability to collect efficacy evidence of a learning tool is constrained by the availability and appropriateness of a measurement tool (e.g., test) to estimate the amount of learning that occurred. If ACT performance and AOP performance were not related, that would indicate a lack of alignment between the two.

Additionally, evidence based on relations to other variables can also be used to support claims about the third and fourth levels of outcomes in Kirkpatrick's training evaluation model – behavior and impact. Recall that behavior is the degree to which learners apply what they learned through the use of the learning tool to real-life events. For example, research has linked higher ACT scores to performance in high school, such as higher GPAs and higher graduation rates. Therefore, we can use validity evidence to make the logical argument that students who improve their ACT performance through the use of AOP will also be more likely to perform well in high school.

We can also use this type of evidence to support the fourth level – impact. Impact is the degree to which targeted outcomes occur as a result of using the learning product. Do individuals who use the learning product have improved outcomes (intended learner outcomes)? As alluded to earlier, the benefit of taking the ACT is not limited to simply receiving an ACT score report. Moreover, the benefit of using AOP is not only to increase one's ACT scores. Rather, higher scores on the ACT are related to many positive, desirable downstream outcomes such as increased chances of acceptance at more selective institutions, higher college enrollment rates, higher college course grades, lower remediation rates, and higher persistence and graduation rates. Therefore, as we did with behavior outcomes, we can use validity evidence to make inferences about impact outcomes; students who improve their ACT performance through the use of AOP will also be more likely to experience the positive college outcomes listed above.

Evidence of Results. Efficacy studies can be conducted to support both behavior and impact outcomes to directly test whether learning improvements gained through AOP transfer to real world contexts (behavior) and improved

intended learner outcomes (impact). For example, do students who use AOP have larger score gains on the ACT than their peers who do not use the product? And do these improved score gains transfer to their current classroom performance as well as translate to downstream consequences of higher college enrollment and completion rates?

Development of ACT's Efficacy Framework

The previous section was hopefully persuasive in convincing readers that an efficacy argument is dependent on the availability and suitability of an assessment to measure that learning. And the suitability of that measurement tool is dependent on the validity evidence supporting that use of that assessment to estimate learning. As such, ACT's Efficacy Framework requires the integration of both a validity argument and an efficacy argument.

Integration of a Validity Argument and an Efficacy Argument into an Efficacy Framework

Intuitively, ACT already knows learning and measurement must go hand in hand. This is evident by a review of our product portfolio. As shown in Table 3, for each ACT assessment solution, there is also a learning solution counterpart. Specifically, at ACT, we not only provide assessments to measure what individuals know and are able to do but also provide learning solutions to improve those KSAOs if individuals need to skill up or if they desire to further propel their learning. For example, for the ACT, we have ACT Academy™, AOP, ACT Kaplan® Online Prep Live (AKOPL), and ACT Recommends – all of which are learning tools to help students better prepare for the ACT. PreACT® is also listed as a learning solution. Even though the PreACT is clearly an assessment, it provides students with a realistic preview of their likely experience on the ACT. In addition, the PreACT score report can provide useful information in terms of areas where students need to improve prior to sitting for the ACT.

Even though implicitly we know the importance of coupling learning and measurement, we do not currently have an Efficacy Framework explicitly describing how to integrate validity and efficacy arguments and evidence into an overarching framework. This is important as we want to accomplish two goals:

1. Develop learning solutions that are most likely to impact the intended outcome.

2. Increase our ability to detect whether a learning solution is achieving its intended outcome through thoughtful study designs, methodology, and data collection.

To this end, we propose an overarching Efficacy Framework, detailing how the alignment between validity evidence of assessments and efficacy evidence of learning products will promote conditions that optimize the likelihood that we will accomplish those two goals.

Table 3. Alignment between ACT's Learning Solutions and Measurement Solutions

Market	Holistic Framework Coverage	Learning Solutions	Assessment Solutions
K-12 (Elementary & Middle School)	Core Academic	Aspire Classroom OpenEd ACT Recommends	Aspire Classroom Aspire Interim Aspire Summative
K-12 (Middle School)	Behavior Skills/SEL	Teacher Playbook	Tessera Middle School
K-12 (Early High School)	Core Academic	OpenEd ACT Recommends	PreACT Aspire Summative/Interim
K-12 (High School)	Core Academic	PreACT ACT Academy AOP AKOPL ACT Recommends	ACT
K-12 (High School)	Behavior Skills/SEL	Teacher Playbook	Tessera High School
Postsecondary	Core Academic	CollegeReady	CollegeReady
Postsecondary	Behavior Skills/SEL	Teacher Playbook	Engage College
High School Through Workforce	Navigation	ACT Profile	Interest and Values Inventories
Workforce	Core Academic	WorkKeys Curriculum	WorkKeys
Workforce	Behavior Skills/SEL	Workforce Playbook	Tessera Workforce

Note. SEL = social and emotional learning

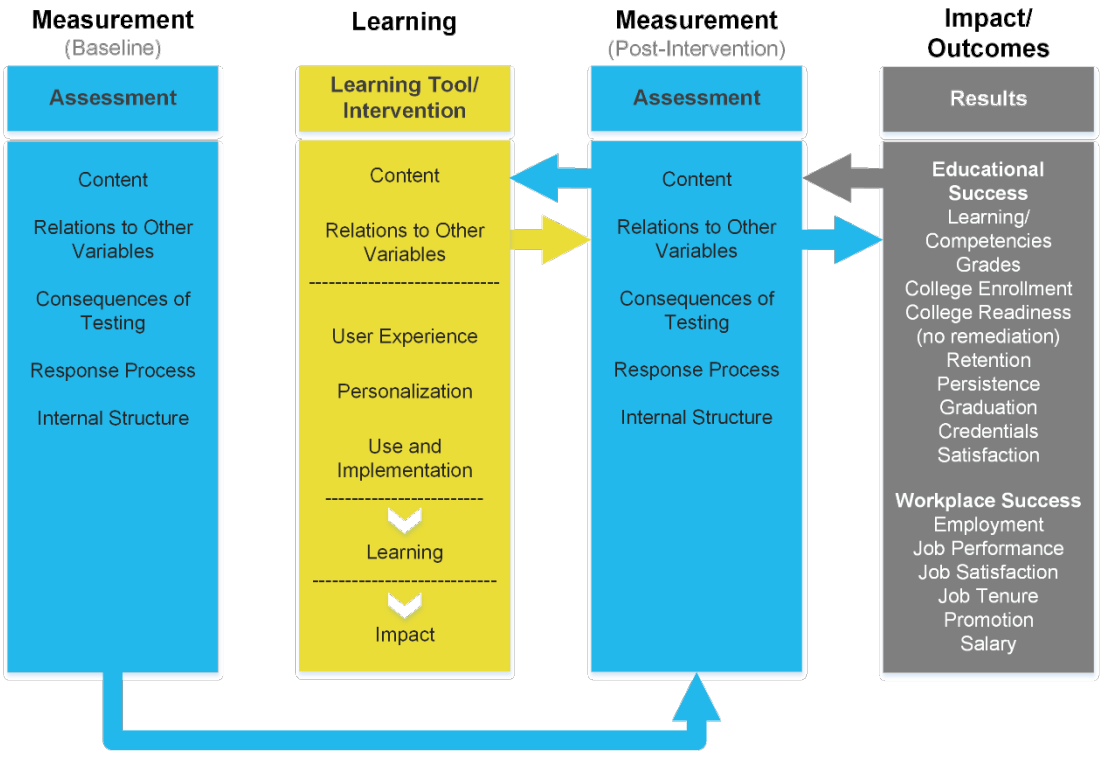
Alignment of Validity Evidence and Efficacy Evidence

Figure 7 illustrates how validity evidence and efficacy evidence work in concert in the development and evaluation of learning tools. As was described earlier, the first step is to identify the learner outcome that we ultimately want to impact. This is represented in the far right column in the diagram. Examples of learner outcomes we may want to focus on for both the educational and workforce settings are provided. These should not be considered an exhaustive list of possible learner outcomes that ACT is interested in improving. As an example, we may want to increase college readiness among all high school students and thus remove the need for remedial education in college. Once a learner outcome is identified, the next step would be to identify the KSAOs necessary for that outcome – in this example, college readiness. We can focus on the ACT, which is a measure of college readiness, for illustrative purposes. To determine what students need to know and be able to do to succeed in college, ACT routinely surveys college faculty to ask them about the prerequisite knowledge and skills incoming students need to succeed in their content domain. This information guides the ACT test blueprint, which specifies which skills will be included for each subject and how they will be measured. This is represented in the diagram by the grey arrow from the Impact/Outcomes column pointing to Content of the assessment in the Measurement (Post-Intervention) column. To empirically test whether the knowledge and skills assessed on the ACT are the skills that are needed to succeed in college, we conduct research showing that ACT scores are predictive of college success, such as course grades. This is represented in the diagram by the blue arrow from the Relations to Other Variables in the Measurement (Post-Intervention) column pointing to the Impact/Outcomes column.

Now that we have documented what students need to know and be able to do to succeed in college, the next step is to develop a solution that helps students improve those exact KSAOs. Therefore, the content covered on the ACT should inform the content that is included in the learning product, which is represented as a blue arrow from the Measurement (Post-Intervention) column pointing to Content of the Learning column. For this example, we will use AOP as the learning solution. To ensure that the content of AOP does in fact align to the content measured on the ACT, we can compare content coverage of AOP to ACT test blueprints (alignment study) or we can examine whether AOP performance is predictive of ACT scores. This is represented in the diagram by the gold arrow from the Relations to Other Variables in the Learning column pointing to the Measurement (Post-Intervention) column.

Within the Learning column, there are two downward-pointing arrows, one of which represents how User Experience, Personalization, and Use and Implementation can impact the degree to which Learning occurs, in addition to the actual learning Content. The other represents how the magnitude of Learning will influence the magnitude of Impact. Finally, the far left column represents the need for measurement of the same KSAOs as the learning solution and post-assessment prior to use of the learning tool to serve as a baseline measure. In some cases, the pre- and post-assessment are the same – in this case, the ACT. Other times, we may propose developmentally relevant, yet still aligned pre- and post-assessment such as PreACT to ACT. This is for similar reasons discussed above indicating the need to link evidence across other columns in the diagram, namely content alignment and predictive validity. Having a baseline measure is useful to estimate individual growth; however, it is not always necessary if students can be randomly assigned to treatment and control conditions and thus should be roughly equivalent on the KSAO of interest. Even when experimental studies are feasible, it is often helpful to have a pre-measure to check how well the random assignment actually worked.

Figure 7. ACT's Efficacy Framework: Alignment of Validity and Efficacy Evidence



The figure above represents a systems approach to learning and measurement, linking evidence across the two to drive the development of efficacious learning solutions as well as inform future enhancements to the product. The framework also guides the different types of research activities and evidence we should collect. In the next section, we propose a framework for determining the level or rigor of evidence needed to guide research activities for specific learning products.

A Framework for Determining the Level of Evidence Needed for Validity and Efficacy Arguments

In previous sections, general types of research activities (refer back to Tables 1 and 2) were identified as providing a basis for different sources of validity and efficacy evidence. In most of the cases, the exact methodology or study design was not specified. Different study designs or methodologies provide evidence that can vary in their level of rigor for supporting different claims. The required level of rigor of evidence should take into consideration several factors when determining the appropriate study design:

1. Use case
2. Stage of the product life cycle
3. Desired strength of validity or efficacy claim and funding
4. Intended audience
5. Return on investment/investment risk

Consideration of Use Case in Determining Needed Evidence: High Versus Low-Stakes Uses

For the ACT assessment, very rigorous evidence is necessary given the high-stakes nature of college admissions. There needs to be clear evidence that the ACT assessment is a valid assessment for that use. This includes ensuring score comparability across test forms and occasions. As mentioned above, ACT forms are equated so that a score earned by one test-taker during the October administration can be compared to the score earned by another test-taker during a different administration. On the other hand, other assessments or learning products are used for more low-stakes purposes such as for diagnostic or formative purposes.[4] For example, Engage® College[5] is used to identify at-risk students once admitted and enrolled on a college campus. Given this use, the rigorous research requirements around developing new items, creating parallel forms, and equating to ensure comparability that are current practice for the ACT is not necessary for Tessera. On the other hand, there may be other issues, such as the fakeability of social and emotional learning (SEL) items, which would require different research activities not required of the ACT.

Consideration of the Product Life Cycle in Determining Needed Evidence: Balancing Innovation and Evidence

In addition to *how* the information gleaned from the assessment will be used, another factor which should also influence the required level of evidence needed is where the product is in its life cycle, in particular for learning products and low-stakes assessments. Requiring too high of a level of evidence early in a product lifecycle can kill an idea before it even has a chance to succeed (Puttick & Ludlow, 2013). On the other hand, requiring too little or no evidence can lead organizations down a costly path in the creation

or acquisition of an undesired, invalid, and/or ineffective product. Therefore, organizations need to wisely balance both a desire for innovation as well as the need/requirement for supporting evidence. To promote these core capabilities, we have researched other intervention frameworks, such as Nesta's level of evidence for interventions framework (Puttick & Ludlow, 2013) and Flay et al.'s (2005) criteria for efficacy evidence, and have aligned them with ACT's Kanban phases. By doing so, we have developed guidelines around the minimum evidence desired and the necessary research activities needed as an idea moves from conception, design, and into the market.

Similar to the Nesta framework, Flay et al. (2005) has provided levels of criteria for efficacy evidence and also incorporates the distinction between efficacy and effectiveness as well as the concept of implementation fidelity in their concept of "ready for dissemination" (refer to Figure 8). These definitions provide useful guides around the type and amount of evidence we should strive to collect for all of our learning tools.

Figure 8. Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination (Adapted from Flay et al., 2005, p. 151)

Efficacy	"Tested in at least two rigorous trials that (1) involved defined samples from defined populations, (2) used psychometrically sound measures and data collection procedures; (3) analyzed their data with rigorous statistical approaches; (4) showed consistent positive effects (without serious iatrogenic effects); and (5) reported at least one significant long-term follow-up."
Effectiveness	"Will not only meet all standards for efficacious interventions, but also will have (1) manuals, appropriate training, and technical support available to allow third parties to adopt and implement the intervention; (2) been evaluated under real-world conditions in studies that included sound measurement of the level of implementation and engagement of the target audience (in both the intervention and control conditions); (3) indicated the practical importance of intervention outcome effects; and (4) clearly demonstrated to whom intervention findings can be generalized."
Ready for Dissemination	"Will not only meet all standards for efficacious and effective interventions, but will also provide (1) evidence of the ability to "go to scale"; (2) clear cost information; and (3) monitoring and evaluation tools so that adopting agencies can monitor or evaluate bow well the intervention works in their settings."

Additionally, we have worked on mapping research activities (not limited to validity and efficacy related activities) to the ACT Kanban process phases. The goal is that these crosswalks will help level set the type and level of research evidence needed as an idea progresses through the Kanban stages and to ensure ACT Research plays an appropriate role across the stages. As indicated above, for a new learning product idea to be submitted to the funnel, one may only provide a logical rationale for why this product is a good idea or present customer feedback. The process does not require a fully functional prototype that has already demonstrated promising results in a multi-site efficacy trial at this stage; though, if that evidence is available, all the better. However, as we progress to later stages in the

Kanban process and more resources are being devoted to the development of this product, stronger evidence is needed to ensure a return on investment. For example, research activities and collecting more evidence in later stages include creating prototypes and documenting validity and efficacy arguments and running validity and efficacy trials. Again, the goal is to balance innovation and evidence to promote the development of new products (innovation) that are mostly likely to improve learner outcomes (efficacy evidence). A similar mindset and set of principles should be applied as we evaluate potential mergers and acquisitions (Puttick & Ludlow, 2012; discussed in more detail below).

Consideration of the Desired Strength of the Claim in Determining Needed Evidence

Mapping the level of evidence needed to the Kanban phases can help determine research activities at each phase. Note that these are initial recommendations and not meant to be prescriptive. Another thing to consider when determining the design of a research study and the associated evidence that will be collected is the strength of the claim that ACT wants to be able to make for specific products. More rigorous research designs result in stronger evidence allowing one to make stronger claims about the product. Pearson's Efficacy Reporting Framework (Pearson, 2018) describes the different study designs that support different types of efficacy claims. We have extended this framework to include anecdotal/qualitative claims as another source, albeit weak, of product efficacy (see Table 4).

Table 4. Strength of Efficacy Claims tied to Study Design

Strength of Claim	Type of Efficacy Claim	Study Designs
Very Weak	Anecdotal/Qualitative Claims	Customer Feedback/ Endorsements Focus Groups
Weak	Descriptive Claims	Surveys Secondary analysis of administrative data Cohort analysis with no controls
Moderate	Correlational/Predictive Claims	Cohort analysis with controls Cohort analysis with no controls
Strong	Comparative Claims	Quasi-experimental Randomized controlled trial
Very Strong	Causal Claims	Randomized controlled trial Propensity score matching Instrumental variables Regression discontinuity

As shown in Table 5, we have labeled different types of efficacy claims and study designs as providing different strengths of evidence, ranging from very weak for anecdotal/qualitative claims to very strong evidence for causal claims based on experimental study designs. We extended Pearson's framework (2018) to include anecdotal/qualitative claims because of the prominent role that customer feedback and product endorsements play in the evaluation of learning products.

With causal claims, we can state that our products are the reason that individuals achieve improved learner outcomes whereas correlational analysis are limited to statements that describe a relationship or association between learning tool use and learner outcomes. Note it is important that the efficacy claims are aligned with the actual study design. If not, the credibility and reputation of an organization is at risk. Having strong efficacy claims is not simply useful for use as marketing collateral for our products but it also has funding implications. If we want states, districts, and schools to adopt our products, a certain level of evidence is necessary. The Every Student Succeed Act (2015) specifies the level of evidence needed to qualify for federal funding dollars, classifying evidence into three categories: strong, moderate, promising. The US Department of Education has stipulated strong parameters in terms of what types of research designs will be considered when including studies in the evaluation of the effectiveness of interventions to be reported in *What Works Clearinghouse* (U.S. DOE, IES, WWC, 2014). Case in point, a review of ACT Aspire was conducted in 2017, and no evaluation could be rendered because no studies of Aspire met the inclusion criteria (U.S. DOE, IES, WWC, 2017):

This intervention report presents findings from a systematic review of ACT Aspire™ conducted using the WWC Procedures and Standards Handbook (version 3.0) and the Transition to College review protocol (version 3.2). No studies of ACT Aspire™ that fall within the scope of the Transition to College review protocol meet What Works Clearinghouse (WWC) group design standards. Because no studies meet WWC group design standards at this time, the WWC is unable to draw any conclusions based on research about the effectiveness or ineffectiveness of ACT Aspire™ on high school and college students. Research that meets WWC design standards is needed to determine the effectiveness or ineffectiveness of this intervention. (p.1)

To be competitive in RFPs and other external funding opportunities, it is imperative that organizations ensure that their research standards are aligned with the criteria of those funding institutions.

Consideration of the Intended Audience in Determining Needed Evidence: Highlighting Evidence that is Most Relevant to Particular Stakeholders

The intended audience is another important consideration when determining the types of validity and efficacy evidence that one may want to collect and highlight. Nichols and Lai (in press) showed that the persuasiveness of different sources of validity evidence varied for different audiences or stakeholders. For example, validity evidence that is most important to psychometricians may not necessarily be the most important or persuasive source of evidence for teachers, parents, or policymakers (Croft, Nichols, & Lai, 2018). Highlighting that other parents endorse an educational app may resonate more with other parents as compared to evidence that there is strong content alignment between the content in the learning app and their child's state standards.

Recall that customer feedback/endorsements were categorized as a very weak claim of efficacy in terms of making causal claims about the effectiveness of the learning product but can still be very impactful in terms of influencing purchasing behavior. To promote the adoption of learning products in students' homes, classrooms, schools, districts, and/or states,

providers should highlight the validity and efficacy evidence that is mostly likely to influence the intended audience's decision to buy the product. That said, it remains important to demonstrate that a learning product is in fact efficacious and is improving the intended learning outcome through rigorous research study designs. To be clear, we are not arguing that this information is not important to collect; rather, the evidence we want to highlight may vary for different stakeholders or audiences.

Consideration of Impact and Impact Risk in Determining Needed Evidence: Mergers and Acquisitions

Finally, research evidence should also play a role in determining where to strategically invest our dollars, not only internally but externally, such as is the case with mergers and acquisitions (M&As). We draw from Puttick and Ludlow's (2012) framework for impact investment to help inform investment decisions arounds M&As. In the framework, investments should be driven by three factors: level of evidence (refer back to Table 5), level of impact (efficacy), and level of impact risk (certainty that the product will result in the stated impact). Level of evidence and impact risk are inversely related as the causal mechanism of impact evidence is unknown when it is based on lower levels of evidence. When the level of evidence is low and thus the impact risk is high, investment decisions around products with a documented negative impact should be avoided.

When the evidence suggests a positive impact, we may want to invest in products with high risk impact, knowing that as part of the investment the expectation is that the impact risk will lower through high quality, rigorous research over time. A caveat of this approach is that an organization needs to be committed to invest additional resources into a product to research its effectiveness after acquisition to ensure the product is having the intended outcome(s) and/or to drive product enhancement(s) based on feedback and evidence as it becomes available. This model should also inform when investment in a product should cease – as evidence accumulates that the product is not effective. Finally, a product that has a positive impact and low impact risk is likely a safe choice and may seem like the best investment decision but it is likely associated with a higher price tag. Organizations need to be able to balance risk and investment dollars, tolerating some level of risk to optimize a likely return on investment.

Conclusion

The number of learning products in the market seems to be increasing daily. Each learning product makes claims – either explicitly or implicitly – that using them will improve one or more intended learner outcomes. Unfortunately, evidence supporting these claims are often lacking. One reason for this may be in part due to the fact that the field of education and education measurement has devoted much less attention to the development of efficacy arguments and frameworks as compared to validity arguments and frameworks. The ideas presented here are intended to support and provide thought leadership and direction to efficacy research being conducted, in general and specifically at ACT. In particular, ACT's Efficacy Framework is intended to guide research activities to ensure that the hallmarks of rigor, science, and scrutiny of our measurement solutions are also built into our learning products with the ultimate goal of ensuring efficacy claims are supported by evidence.

Notes

1. In our Efficacy Framework, we make the distinction between efficacy and effectiveness and highlight the importance of researching implementation or contextual factors that may bolster or undermine the impact of a learning product on intended learner outcomes.
2. Learning can be further subdivided into types of learning goals or levels of learning. A popular taxonomy of learning, Bloom's Taxonomy, has identified six levels of learning: remember, understand, apply, analyze, evaluate, and create to reflect more sophisticated cognitive processes (Krathwohl, 2002). Moreover, different types of instructional practices are more or less effective depending on the learning goal or knowledge component (Koedinger et al., 2012).
3. A/B testing uses an experimental design where two or more version of a webpage or screenshot are presented to users at random; statistical tests are performed to understand which version is performing better in relation to a desired outcome (e.g., higher engagement, improved learning).
4. Formative solutions consist of resource, non-scored assessment, and professional development tools that support learning in the classroom or are embedded in other learning processes. This is another mechanism through which learning is verified via measurement.
5. Social and Emotional Learning (SEL) looks beyond core academic skills to provide a more comprehensive picture of the person, providing augmented views to summative cognitive assessments in education and work contexts. SEL examines a consequential set of success factors and personal skills impacting student and employee success that are amenable to change and development. At ACT, SEL spans both the learning and measurement circles in the Venn diagram in that we want to both help individuals develop SEL skills (e.g., Teacher Playbook) as well as document whether students have developed those skills (e.g., Tessera).

References

ACT. (2016). *ACT National Curriculum Survey 2016*. Iowa City, IA: ACT.

Allen, J. (2013). *Updating the ACT College Readiness Benchmarks*. Iowa City, IA: ACT.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bjork, E. L., & Bjork, R. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacjer, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology in the real world: Essays illustrating fundamental contributions to society* (2nd ed., pp. 59-68). New York, NY: Worth.

Chen, H. T. (2015). Logic Models and the Action Model/Change Model Schema (Program Theory). In H. T. Chen (Ed.), *Practical Program Evaluation: Theory-Driven Evaluation and the Integrated Evaluation Perspective* (2nd ed., pp. 58-93). Thousand Oaks, CA: SAGE Publications

Croft, M., Nichols, P., & Lai, E. (2018). *In the eye of the beholder: Stakeholder perceptions of validity evidence*. Paper session presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Pearson. (2018). *Efficacy reporting framework*. London, UK: Pearson, Inc.

Every Student Succeeds Act (2015). S. 1177; Pub.L. 114-95.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., ... & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention science, 6*(3), 151-175.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of the American Society of Training Directors, 13*, 3–9.

Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resource development* (2nd ed., pp. 301–319). New York, NY: McGraw-Hill.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science, 36*(5), 757-798.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218.

Nichols, P. D., & Lai, E. (in press). *Role of stakeholders in validity*. Iowa City, IA: ACT.

Pellegrino, J. W., & McCallum, W. (2017). Conceptualizing and measuring progress toward college and career readiness in mathematics. In K. L. McClarty, K. D. Mattern, & M. N. Gaertner (Eds.), *Preparing students for college and careers* (pp. 23-34). New York, NY: Routledge.

Praslova, L. (2010). Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in higher education. *Educational Assessment, Evaluation and Accountability, 22*(3), 215-225.

Puttick, R., & Ludlow, J. (2013). *Standards of evidence: An approach that balances the need for evidence with innovation*. London, EC: Nesta.

Puttick, R., & Ludlow, J. (2012). *Standards of evidence for impact investing.* London, EC: Nesta.

US Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2017, May). *Transition to College intervention report: ACT Aspire*™. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_aspire_053117.pdf.

US Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2014). *Procedures and standards handbook* (Version 3.0). Washington, DC: US Department of Education.

ACT is an independent, nonprofit organization that provides assessment, research, information, and program management services in the broad areas of education and workforce development. Each year, we serve millions of people in high schools, colleges, professional associations, businesses, and government agencies, nationally and internationally. Though designed to meet a wide array of needs, all ACT programs and services have one guiding purpose—helping people achieve education and workplace success.

ACT®

ACT.org/research