# Smart Items™

**Dr. David Foster**
Caveon Test Security

## Stop Test Fraud
## Improve Fairness &
## Upgrade the Way You Test

caveon™

# SmartItems™

Stop Test Fraud

Improve Fairness &

Upgrade the Way You Test

**Dr. David Foster**

Caveon, LLC

© 2020, Caveon, LLC. All rights reserved. Do not copy or distribute without permission.

# TABLE OF CONTENTS


00 **Preface & Introduction**
*A Little Background & The Future with SmartItems* page 1

01 **The Critical Needs of the Assessment Industry**
*Harnessing Technology to Solve the Problems Facing Our Industry* page 9

02 **What Makes an Item "Smart"?**
*The Nine Properties of a SmartItem* page 23

03 **The Difficulty Question**
*Logical, Theoretical, Practical, & Empirical Arguments for SmartItems* page 39

04 **Errors, Errors, Everywhere**
*Dealing with Common, Serious, Systematic Errors in Measurement* page 55

05 **Implementing SmartItems**
*The Technology Tools You Need to Get Started* page 76

06 **Creating and Evaluating SmartItems**
*Design, Implement, and Evaluate* page 86

07 **Evidence from Case Studies**
*Early Adopters and the SmartItem Impact* page 109

08 **Evidence from Scientific Experiments (Part 1)**
*Do SmartItems Work? We've got the answer.* page 121

09 **Evidence from Scientific Experiments (Part 2)**
*Do SmartItems Work? We've got the answer.* page 136

10 **Evidence from Data Simulations**
*Demonstrating Difficulty Equivalence* page 156

00 **Conclusion, References, Acknowledgements, Resources, and Glossary**
*The Time is Right for Embracing SmartItems* page 163



# ABOUT THE AUTHOR

## David Foster

Ph.D. | CEO & President of Caveon

A psychologist and psychometrician, David has spent 36 years in the measurement industry. David's rich background in research and innovation has motivated him to apply a scientific approach to addressing problems in testing, mingling a healthy skepticism of current best practices with a research-based and objective appraisal of new solutions.

David is credited with introducing computerized adaptive testing and simulation-based performance testing as part of Novell's IT pioneering certification program in the early 1990s. Following his tenure with Novell, he co-founded Galton Technologies, which was acquired by Thomson Learning in 2000.

In 2003, David co-founded the industry's first test security company, Caveon. Under David's guidance, Caveon has created new security tools, analyses, and services that protect clients' exams throughout their entire lifecycle. During the past decade, amid rising concerns about fairness in testing, David has focused on changing the design of items and tests to reduce the debilitating effects of cheating and testwiseness.

David holds a Ph.D. in Experimental Psychology from Brigham Young University. He completed his post-doctoral fellowship in Psychobiology at Florida State University and has served on numerous boards and committees in the accreditation, testing, and measurement industries. David is a past President of the Association of Test Publishers (ATP), and has served on the certification accreditation committee and Board of Directors for the American National Standards Institute (ANSI) and on the council and Board of Directors for the International Test Commission (ITC). In 2002, David founded the Performance Testing Council to help raise awareness of the principles required for quality skill measurement. He has authored numerous articles for industry publications and journals and has presented extensively at conferences.

David's outlook and enthusiasm have led to the creation and implementation of a number of innovations during his career, including the first large-scale global use of computerized adaptive testing, the effective use of simulations in tests, the Discrete Option Multiple Choice (for which he owns the US Patent), the introduction of online proctoring, and now SmartItems. After 36 years in the testing industry, David shows no evidence of slowing down, and he, no doubt, has a few more contributions to make.

*It is time for us to embrace innovation and technology and bring the multiple-choice question into the 21st century, where it can evolve to address our current needs, and proactively tackle the problems of the future.*

~ David Foster

*SmartItems are invulnerable to cheating and theft, relieving SailPoint of the same security concerns that have plagued every other IT certification program for the past three decades.*

~ Arthur Altman

Program Manager
Global Education Services
SailPoint Technology Holdings

PREFACE

# THE FUTURE WITH SMARTITEMS

Imagine a world where cheating on high-stakes tests is virtually impossible.

Imagine a world where test scores are entirely free from the influence of test fraud—a fairer world where cheaters are unable to prosper at the expense of well-deserving test-takers.

Imagine a world where a hiring manager can trust test scores from applicants without reservation, a world where students know they can confidently succeed in college, a world where individuals from other cultures or individuals who have disabilities can demonstrate their skills on a level playing field.

Imagine a world where teaching to the test is encouraged, because the test itself is finally able to measure what it should.

Imagine a world where the blatant theft of test questions, while against the rules, does not compromise or impact the test. This is a world where exams last forever, and where millions of dollars are not wasted on creating dozens of test forms or on the constant production of replacement items.

Imagine a world where tests can no longer be "gamed," where scores are not inflated because some individuals are better at taking tests. This is a world where there are no shortcuts to success, and no arbitrary barriers, either.

Imagine a world where each person learns at their own pace, where progress and mastery are demonstrated using tests that are fair, protected, convenient, low-cost, accessible, repeatable, and where the uses of test scores are valid.

Imagine a world where the efforts of those in the testing industry are praised rather than criticized, a world where the value of testing is proven every day. It is a world where our industry is viewed with respect by students, teachers, parents, school counselors, professors, companies, governments, and media representatives due to our success in solving long-standing problems, such as test fraud, testwiseness, inconvenience, inaccessibility, and others.

Imagine a world where we have shed our reliance on outdated methods of testing, and where we find ourselves in sync with the fast-paced technology that drives us toward progress. It is a world where we anticipate and even guide change, rather than getting dragged along and surprised at every unexpected turn.

Imagine a world where costs to testing programs decrease, where the entire testing process—

from design and development to test administration and scoring—is affordable to all, including students, parents, workers, companies, and educational institutions.

It is difficult to envision this imagined world. All of your tradition and training and advisors will shout that "it isn't possible," that "it is too good to be true." You will think and hear dozens of reasons why this world can never be. You may end up concluding it isn't worthwhile to even try.

But imagination has a way of becoming a reality. And that reality may be closer than we think.

## Introducing SmartItems.

A SmartItem is a ground-breaking invention that will disrupt the world of testing as we now know it. SmartItems are still test items, just like smartphones are still phones. However, like smartphones, they impact our world in a myriad of incredible ways. SmartItems can lead the way to our imagined world. They can show us what's possible and encourage the development of other beneficial innovations.

Like all new things, SmartItems are not risk-free. However, the risk is insubstantial when compared to all the benefits they offer. The discomfort over doing something new and unusual will be momentary, quickly swallowed up in the excitement of solving the long-standing problems our industry currently faces.

**FOR A GROUNDBREAKING SOLUTION THAT WILL CHANGE THE WORLD OF TESTING FOR THE BETTER ... READ ON.**

# A LITTLE BACKGROUND



In 1982, I was abruptly introduced to the world of testing. More importantly, I was introduced to the technology of testing. A Utah company, WICAT Systems, Inc., received funding from a large corporation, and with the support of a major university, created a mainframe-based, K-12 computer-assisted learning system. It was designed to include a way to assess students, but that piece was still on the drawing board. At the time, I was attending Western Carolina University and looking for a research position at a university. I had just completed a post-doctoral fellowship at Florida State University in, of all things, the chemical senses of taste and smell. I was trained as an experimental psychologist, a scientist. I had conducted and published experimental research with pigeons, humans, rats, dogs, and chicks. Jobs were scarce, so I was intrigued when a good friend from Brigham Young University asked if I could help with the assessment component of this new educational technology. While I was trained in concepts of scientific measurement, and knew next to nothing about testing, I accepted the position.

It took me some time (and no small amount of frantic study) to learn about the theory and practice of testing, as well as about the particular challenges facing the field in 1982. At that time, computerized testing was little more than a pioneering effort, albeit an exciting one. Computers had the potential to create incredible new tests, and building computerized adaptive tests for all grades and

**"SMARTITEMS COULD BE LIKE THE DRIVERLESS CAR, THE SPACESHIP TO MARS, OR THE SMALLPOX VACCINE. IT COULD BE A TECHNOLOGY THAT TRIGGERS INNOVATIONS YET UNIMAGINED, INSPIRES US, AND TEACHES US TO MASTER THE FUTURE."**



subjects was one of my team's first goals. We were advised in our efforts by some excellent consultants, including Ron Hambleton (University of Massachusetts, Amherst), Anthony Nitko (University of Pittsburgh), and Ron Berk (Johns Hopkins University). With the help of these experts, we were able to create a comprehensive suite of computerized exams.

For the nearly four decades since I received that transformational job offer, I have been involved with creating computerized exams and the systems to support them. Looking back over that time, I see that a primary thread connecting my work has been figuring out new solutions to the problems I encountered. Some of those problems had existed for a long time, while others were only introduced as computerized testing expanded. I was determined to address and fix these problems, particularly those that dealt with sources of unfairness or issues with making tests more relevant. I am the first to admit my efforts were rarely guided by "best practices." I have always believed technology is our most powerful influence, and unfortunately, best practices tend to be in short supply when working at the forefront of cutting-edge technology.

**REAL CAREER EXAMPLES:**

## HARNESSING TECHNOLOGY

**1**

While I worked at Novell, I was asked to create exams that measured the performance of candidates on the NetWare operating systems. Traditionally, the exams simply asked questions about NetWare, and our certification credential got hammered in trade journals for not producing candidates who could actually utilize and applythe software properly. We embarked on a plan to create simulation-based exams and built a new exam in just a few months. During the test, examinees were given questions that required the use of a NetWare simulation. With this new model, the concerns about the capability of candidates who passed the exam disappeared overnight.

**2**

I had another opportunity to improve testing through technology when I was asked to address an issue of fairness involving translated exams. On our certification exams, technical networking terms such as "spooler" were sometimes translated into other languages. While all candidates understood the English term, few understood the translated word, even if it was in their native language. We worked on the translation process, but the breakthrough came when we changed the design of the test to include an ENGLISH button on the item screen. If a translated term was encountered and not understood, clicking on the ENGLISH button brought up the item in the original language. This solved the problem, addressed complaints, and eliminated any unfairness.

My experience with technology is best described this way: If applied carefully and logically, innovations can fit naturally into our traditional procedures and quickly provide solutions to the problems we all face. Whenever I determined the status quo was unacceptable, I trusted innovative technology as a solution, and was rarely disappointed. Follow-up research almost always confirmed that we addressed the problems and improved the quality of the exams. Otherwise, the research pointed to new ways to improve our initial attempts.

This storyline has been repeated many times throughout my career: 1) recognize a problem, 2) use technology, and 3) experience a successful outcome. My efforts were rewarded often enough for me to trust innovation going forward and to remain optimistic about the future of the field.

The success of this formula is the reason I so earnestly trust in the application of computing technology to confront the problems facing the testing industry, even before the appearance of best practices and confirming research. If we want to keep up with the modern pace and meet today's needs, I honestly believe we can't wait
for an abundance of research or experts' approval before opening our arms and embracing this formula.

This book is about SmartItems, the most recent technological innovation I have introduced. SmartItems were born out of a severe need for the improvement of test security and for the elimination of resulting unfairness. Our current practices are proving to be ineffective. SmartItems neuter cheaters and thieves by literally preventing almost every improper and illegal method they use to inflate test scores unfairly. I know this statement is astonishing—almost unbelievable—but I stand by it. How SmartItems achieve this goal, and their myriad of other benefits, are discussed throughout this book.

I will use both theoretical and empirical evidence to help the reader understand not just the benefits, but the necessity of using SmartItems. Classical and modern test theories support the use of such an unusual question type, and empirical studies (including scientific experiments, case studies, and simulations) demonstrate its value and support its use. Beyond evidence-based reasoning, however, I make a call to logic—a call for all to understand why we must embrace technologies in



Recognize a Problem



Embrace Technology



Experience Success

**"THE WORLD OF TECHNOLOGY IS FAST-PACED AND EVER-CHANGING... WE CAN'T WAIT FOR OTHERS' APPROVAL BEFORE OPENING OUR ARMS AND EMBRACING IT."**

testing—technologies like SmartItems.

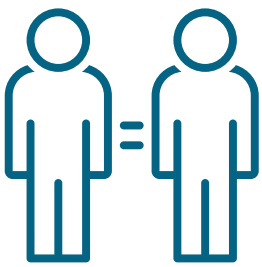
We in the field of testing or assessment are getting hammered mercilessly by real problems, including our inability to deal with test fraud. When faced with a choice of continuing down that path or aggressively adopting a technology that stops fraud in its tracks, the choice is easy.

SmartItems offer exciting improvements for testing professionals. This technology will remove serious obstacles that currently limit our field and will raise our industry to a new level of relevance in society. Accepting—and embracing—technology is the first step in figuring out how we in the testing industry can harness the power of technology at technology's pace, rather than at our traditional stride.

The field of testing came to my rescue in the early 1980s. It provided me a career and a purpose. I love it, and I am indebted to it. From my very first day, I've made changes and introduced new concepts, many of which are now widely used.

However, I can say with absolute confidence that SmartItems have the most promise out of any technological innovation. In our industry, SmartItems are like the driverless car, a spaceship to Mars, or the smallpox vaccine. It could be a technology that triggers innovations yet unimagined, inspires us to explore new depths, and teaches us to master our future.

After nearly four decades in this industry, I am astounded some programs still administer paper-and-pencil tests, traditional multiple-choice questions continue to dominate, cheating is still a great way to get a high score, tests are still unfair to many populations, and costs of test development and test administration remain high.



### The Concept of "Standardization"

Standardization is our attempt to keep items, tests, scoring, test administration, and the rest of the testing experience as similar as possible for all test-takers. The logic of standardization is this: if we hold as much of testing as possible static, the test scores will be fair and comparable. But here's the rub: our efforts at standardization almost always fall short. By trying to keep things the same, we unwittingly contribute to the success of thieves and cheaters. We make their task easier. To understand and accept SmartItems, readers will need to relax their emotional hold on strict standardization. In doing so, they will come to realize that random change, when purposefully incorporated into tests, can be a good thing.

None of these limitations are acceptable. None need to exist today.

Nothing has been more rewarding in my testing career than being able to solve problems, large and small, with technology. For whatever reason, I'm a person who accepts the risks associated with challenging the tenets of traditional approaches. From the first foray, I expect the end result will be positive. Luckily for me, this instinct is usually rewarded. When embraced, technology has improved reliability, improved validity, managed security, reduced costs, and more. With that history, who can blame me for an optimistic commitment to the use of technology in testing? If we can unleash inventions such as SmartItems, the future of this industry will be unbelievably exciting.

The book is designed to make SmartItems understandable to parents, students, workers, prospective workers, lawyers, government officials, testing experts, scientists, teachers, psychometricians, and any other group with a vested interested in the valid use of test scores. In other words, I tried to make this book accessible to a broad audience. As such, I have limited my use of technical language and included more examples, stories, and visuals. Where technical descriptions were unavoidable, I gave brief explanations. Hopefully, the result is an engaging book that informs and enlightens all interested persons while satisfying the appetites of testing experts.

I hope you read this book with an open mind about how testing can and must change. I hope you have questions, and that you'll be patient as the answers unfold—or maybe even willing to help find the answers. I hope you also look forward to a fun, challenging, changing, and fast-paced world of testing.

View
**SmartItems in Action**

http://bit.ly/2EgSmt



Chapter

# 01 THE CRITICAL NEEDS OF THE ASSESSMENT INDUSTRY

## HARNESSING TECHNOLOGY TO SOLVE FOUR PROBLEMS FACING OUR INDUSTRY

Necessity is the mother of invention. As I see it, the testing industry is currently plagued by four critical problems that need to be addressed. While these needs have afflicted our industry[1] for decades, the consequences of not tackling them have become dire in recent years. It's time we harness the power of technology to deal with them.

### Need #1 | Test Security

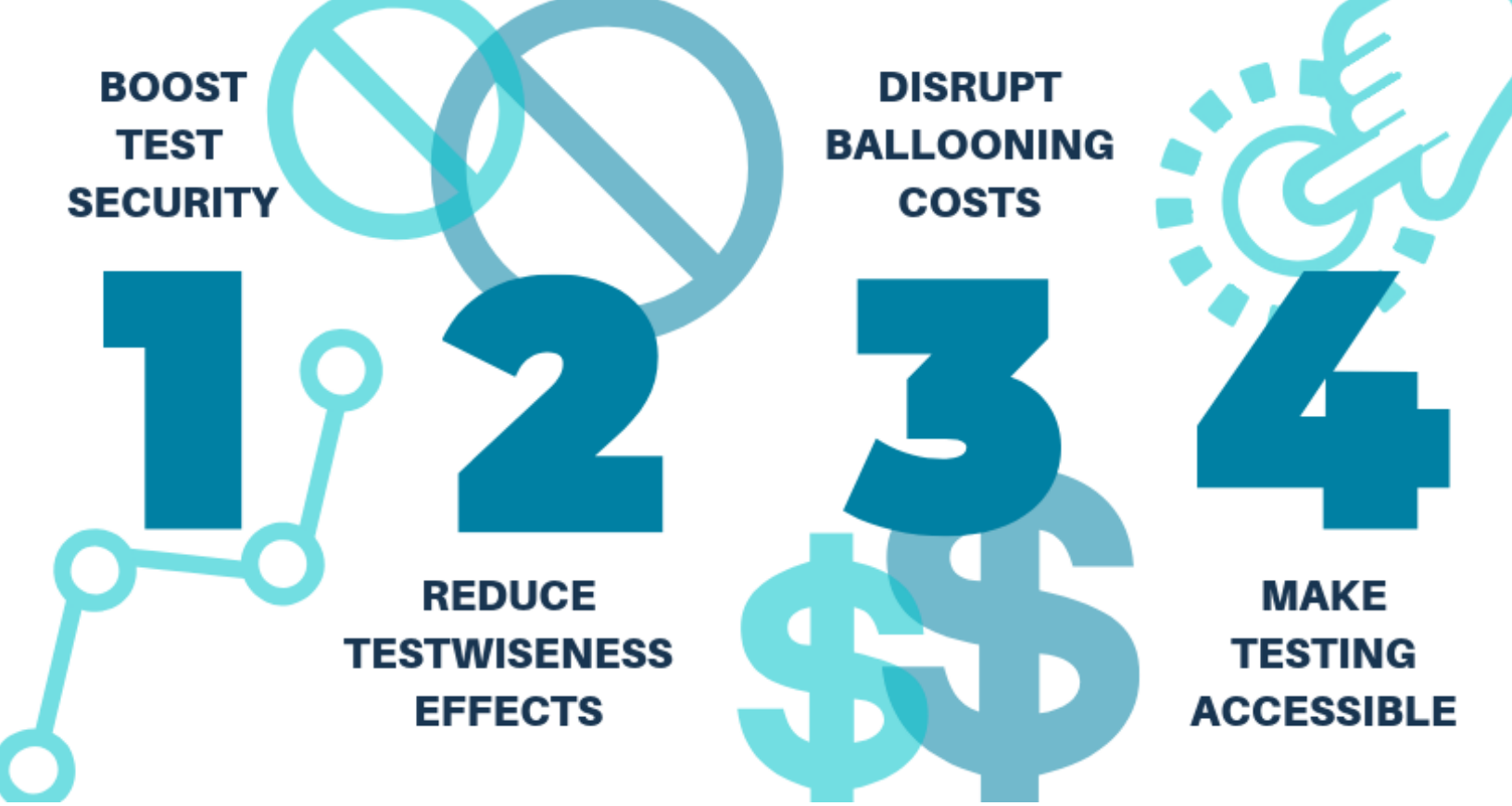
There is not a single high-stakes exam anywhere in the world that is not currently at risk from cheating and theft; all-too-common public scandals attest to our ineffectiveness when it comes to securing exams. The unfortunate truth is most of the test fraud occurring every day probably goes undetected and unreported. Sadly, the risks facing our industry are only growing. New technologies have increased the ease and frequency of test fraud and compounded the consequences of weak test security.

### Need #2 | Testwiseness

The traditional (and most popular) method for the design of tests uses the multiple-choice question type. This introduces a significant source of unfairness by allowing certain individuals to use testwiseness, exploiting unintended testing cues to receive a higher score. An example of this is comparing multiple-choice options and eliminating the least likely options. Testwisesness allows some test takers to benefit by using skills unrelated to the content of the test, and occurs at the expense of other test takers who do not possess those same skills. Allowing this major source of unfairness to continue goes against both our industry standards for quality measurement and our ethical principles.

### Need #3 | Ballooning Costs

As anyone working in this industry knows, developing and administering tests is an expensive endeavor. These costs increase as




security breaches force us to rewrite and republish quality exams. This financial burden is becoming untenable for many high-stakes test publishers and their partners and customers.

### Need #4 | Inconvenience

The process of registering and taking a test is undeniably inconvenient for test takers around the world. We live in a world where technology has brought banking and doctor's appointments into the comfort of our own homes, yet the process of taking a test remains difficult and bothersome.

For high-stakes testing, it is typical for a person to go to a testing location to take an exam. From registering, scheduling, driving to a testing center, and waiting weeks for results, it is a lengthy and challenging process. In addition, individuals who live far away from a testing center who are unable to find childcare or unable to take time off work, find it impossible to take important and impactful exams.

The testing industry is plagued by recurring problems including test security, unfairness from testwiseness, costly test administration, inconvenience, etc. This, however, is not a comprehensive list. Other issues include accessibility by populations with mental or physical disabilities, exams that are too long, exams that don't provide enough time, scoring and reporting issues, cultural and language adaptation issues, international reach limitations, etc.

For each of these problems, technology offers solutions. Lucky for us, the technology is both available and accessible. Our capability of accomplishing and solving many of the problems facing our industry is obtainable—and within arm's reach. The utilization of SmartItem technology alone could solve many of the critical problems listed above.

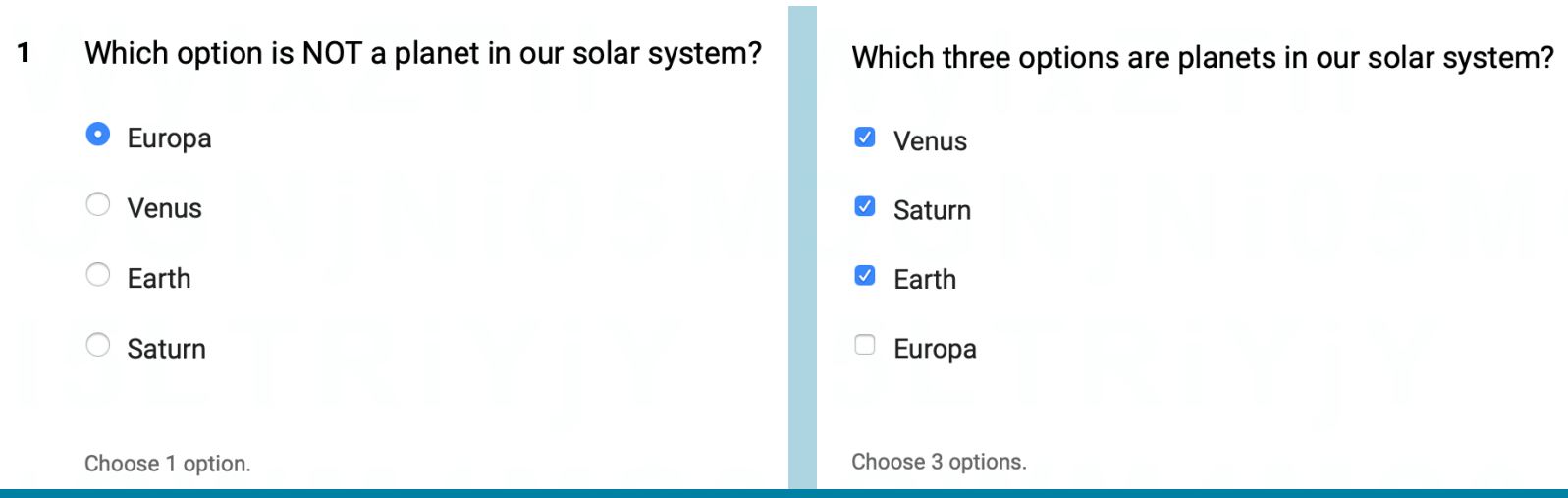
## It's All About Validity!

It is essential you understand the fundamental reason why we need to address the four problems listed above. In the field of testing, our most valuable asset is the validity of our test scores. Validity can be rephrased in less technical language as "the ability for individuals to make good decisions based on the test scores we provide as an industry."

Establishing validity is the cornerstone of quality testing and results in scores people can trust. I cannot emphasize this enough—validity is, by far, the most important criterion for any test. Does the test measure the skills we are interested in, and is it measuring these skills well? Is there evidence that test scores can be interpreted correctly for the intended purposes of the test? A lack of test security, unfairness from testwiseness, costs considerations, and inconvenience all negatively impact validity.

Having been a part of the world of testing for a long time now, I've seen and experienced the entire range of testing problems that threaten validity. One of the first validity challenges I encountered was while I was in charge of creating certification exams for Novell, a software company in the early 1990s. At the time, I was querying what to do about "negatively worded" multiple-choice question—the type of question where the word "NOT" is inserted in the stem.[2] At the time, negatively worded questions like this often existed for no better reason than to make it so that the candidate need only select a single option. Unfortunately, the word "not" made it difficult to create items that were easy to understand. This was especially true for individuals who were unfamiliar with the multiple-choice format, or who were taking the test in a second language. Even with candidates from the United States who were familiar with the format, these backwardly worded stems were challenging to process and seemed unnecessarily tricky.

Having access to one of the first commercially used computerized tests, we were able to come up with a solution to this problem quickly. Why not just ask the test taker to select the several correct options rather than the single incorrect one? We would administer a "choose two" or "choose three" question to the candidates, using checkboxes rather than radio buttons to allow the selection of more than one option. We expected this question format would be easier for the candidate to understand and answer. Subsequent research supported this assumption. Even 30 years ago, we were able to use technology with relative ease to address and eliminate sources of invalidity.



The issue of negatively worded stems was a significant, yet small, validity problem compared to the critical problems of test fraud, testwiseness, costliness, and inconvenience. Critics of testing, often media reporters, have correctly recognized these problems (albeit sometimes misdefining and misdescribing them) and have splashed them across the front pages of prominent newspapers and on the Internet. These public criticisms hurt, yet they are often accurate.

Unfortunately, during my time in this industry, I've rarely seen the testing community publicly (or even privately) acknowledge the truth in these criticisms, and certainly, have never seen an organized industry-response to them. We continue to develop and administer tests using century-old methods, only reluctantly utilizing technological advancements. Forces of economics and misguided concerns about standardization encourage major players to maintain a glacial pace of change. This reluctance to change takes place in both our daily work and in the way we train future practitioners in this industry.

## AUTHOR'S NOTE:
## THE RATE OF TECHNOLOGICAL CHANGE

On a personal note, it has been more than four decades since I took undergraduate psychology courses on my way to a degree in psychology. Recently, I took the opportunity to review the course syllabi and textbooks current psychology students use. I was astounded to discover very little has changed. Students would get just as much out of those courses now if they were using my textbook from 40 years ago. It's no wonder training young students in the science and art of testing feels stagnant despite the crucial innovations that have been introduced.

For example, I recently did a Google search using the phrase "how to write a test item." I found that universities hosted nine out of the top ten websites. These sites described the steps and rules for writing a multiple-choice question, but they included several rules that are no longer relevant. The "training materials" universally stated a multiple-choice question must have a single correct option. However, for the past three decades, there have been strong multiple-choice question alternatives that routinely use more than one correct option. (Sireci & Zenisky, 2006; Cronbach, 1941).

This stagnation is worrisome. Given the pace of technological change worldwide, any industry that changes at a pace slower than the useful technology available will fall behind, becoming more and more irrelevant. It is particularly damaging when the education of young student testing professionals changes infrequently, or not at all.

I've had discussions with colleagues who are concerned that our industry is becoming irrelevant, which may make us replaceable. I share their fears. We are not adapting fast enough.








## 40 Years of Technology in Testing

Computers were introduced in a large-scale way to the testing world in the early 1980s, with personal computers quickly replacing mainframes. Since that time, the use of technology in testing has remained primarily superficial. It is considered "best practice" for a computerized version of a test to mimic the old paper-and-pencil world instead of embracing a new range of capabilities.

### Consider these examples of surface-level technological changes:

- When converting items from paper-based tests to computer-based tests, items are made to look and act the same on computer screens as they always have on paper.
- Nearly all computerized tests allow a test taker to review previously seen items to review and change answers. This was a natural part of taking paper-based tests, but is not necessary when taking tests on computers.
- In most educational settings, students generally take tests as a group in computer labs or similar arrangements. This is patterned after the use of auditoriums when paper tests are administered.
- The use of in-person proctors for observing test takers is still considered the "gold standard" for proctoring, despite the value of online proctoring using webcams, biometrics, computers, and the Internet.
- The myriad of new item types made possible by the computerization of tests in the 1980s represent only a tiny proportion of items used in both high-stakes and no-stakes tests. Tests today continue to be dominated by the traditional multiple-choice question.

Now, consider these examples for areas where simple and readily available technological fixes could have positively changed the way we give tests, but have yet to be embraced:

- Testing programs predominately use government-issued and school-issued ID documentation to allow examinees to take tests.
- Many high-stakes testing programs in the certification, education, and employment testing realms still use paper-based tests.
- Testing programs continue to rely on proctors, both online and in-person, as the keystone of test security.

Finally, consider these areas where the testing industry has embraced technological innovations, resulting in significant benefits:

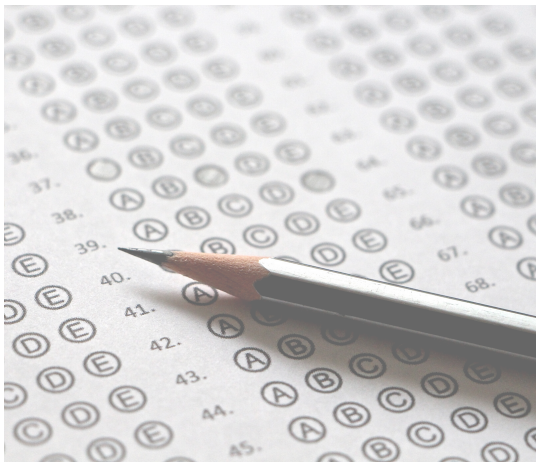
- The use of scanning/scoring technology beginning in the middle of the 20th century greatly improved the ability to process and score millions of educational exams.
- Computerized adaptive testing (CAT)[3], which has been used since the 1980s), provides better security, improves test taker motivation, and shortens exams.
- The Internet is used to quickly and inexpensively transfer tests and test results to and from any part of the globe.
- The Internet is used to "stream" tests item-by-item to a test taker's location, without the delays and security risks of downloading exams to local servers.
- The increase in the storage capacity of technology systems has allowed the use of more complex exams, as well as the ability to collect more detail on test-taking behavior.

These innovations have fundamentally changed our methods of testing and have brought about substantial benefits. Unfortunately, this list remains short, despite extensive opportunities to expand it.

## Suggestions for Technology-Based Solutions

We started this chapter by learning there are four critical problems facing the testing industry that desperately need solutions. We then analyzed the ways technology has (and has not) already changed the world. We will now look at recommendations for how the innovative technology of SmartItems can effectively address the crippling impact of test fraud, testwiseness, costs, and inconvenience.







*It is by preventing, rather than detecting, test fraud that we can once and for all address the pressing need for better test security. This is the most important objective all high-stakes testing programs can pursue today.*

## How SmartItems Improve Test Security

It's reasonable to suggest the current lack of effective test security presents the greatest threat to the validity and trustworthiness of exam scores. The real problem with successful cheating is that it makes it virtually impossible to tell the difference between a valid score from one obtained by cheating. Cheaters appear to be top-performers, and honest examinees are hurt as a result.

If a set of scores is canceled because cheating is suspected or confirmed, the scores of the non-cheaters are often canceled as well. This occurs because today's methods for detecting cheating at the individual test taker level are less precise. It is more practical (and often the only option) to cancel all the test scores in a particular geographical area or on specific days of a testing window.

For example, in 2016, ACT canceled all test scores in South Korea and Hong Kong, affecting about 5,500 students. ACT spokesperson, Ed Colby, admitted the action impacted "innocent students who had no involvement in any kind of wrong activities." (Reuters, 2016) The actions of a few unethical individuals are devastating to those honest individuals who prepare well and try hard. Even if the scores of honest test takers are not canceled, a few individuals caught cheating can call into question the reputability of the scores of every person who took the test. If a handful of individuals are caught cheating on a certification exam, everyone else who received the certification honestly will face doubt and scrutiny.

**The desperate need for an effective test security solution should be clear. Our current approaches are simply not good enough.**

We must analyze what more we can do to prevent the unfairness that results from cheating. The obvious response is to say we should simply get more surgical with our ability to detect cheating. If we can accurately isolate just the cheating individuals, we can cancel only their test scores and protect the validity of the remaining tests. Unfortunately, cheating is designed to be well-hidden and difficult to detect. Some methods for cheating, such as having access to the exact questions on the test in advance (also called pre-knowledge), are impossible for proctors to detect during test administration. Creating methods for detecting cheating with 100% accuracy is a tall, if not impossible, order.

Where does this leave us? If cheating and test theft are difficult to detect, we must turn our attention to preventing cheating from happening in the first place—specifically in preventing cheating through pre-knowledge. Let me repeat: we must stop cheating from happening in the first place.

It is by preventing, rather than detecting, test fraud that we can once and for all address the pressing need for better test security. This is the most important objective all high-stakes[4] testing programs can pursue today.

One effective solution is SmartItems. SmartItems eliminate the debilitating influences of test theft and mitigate most forms of cheating. I first introduced the SmartItem method and technology to the testing industry in the spring of 2017 at the Innovations in Testing Conference put on by the Association of Test Publishers. In a session titled "The Future of Multiple Choice" (and discussed in the webinar referenced below titled "Using Innovative Item Design to Make Your Life Easier"), I described SmartItems as a possible replacement for traditional multiple-choice. I did this by first providing evidence that the venerable old item type has outlived its usefulness, and may even be doing more harm than good.

In the presentation, I covered the fundamentals of SmartItems, discussing what exactly a SmartItem is, why it is such an important innovation, and how it is different from other security solutions. All of these topics will be covered in detail in the following chapters.

In the meantime, understand that SmartItems render tests that are entirely unpredictable while still measuring the desired skill. The questions on the test are selected from a random sample from the universe of all possible items for the test. For those unfamiliar with the theory of testing, tests are supposed to be a random sample of all possible items, but are usually a set of fixed items. For those of you who work in this field, this claim for SmartItems might sound too good to be true. But SmartItems are uniquely designed to accomplish this goal. In meeting that goal, they make the theft of items functionally impossible and prevent most forms of cheating as well.

**SmartItems:**
Using Innovative Item Design to Make Your Life Easier

**DOWNLOAD THE PRESENTATION HERE**





*It is time for us to take a stand against allowing an irrelevant skill like testwiseness to have such a significant impact on test scores.*

## How SmartItems Reduce Testwiseness Effects

If test security threats are the most crucial source of invalidity, a close second is the long-standing issue of unfairness resulting from testwiseness. Initially identified as an issue by Thorndike (1951) nearly 70 years ago, testwiseness refers to the ability of some people to use cues inherent in multiple-choice questions, to "game the test," and to get a higher score. Testwiseness means a test score is not just representative of the test taker's capability in or knowledge of the skill being tested, but is enhanced because of irrelevant skills that are facilitated by the multiple-choice format.

Testwiseness directly and unfairly affects the validity and trustworthiness of our test scores, and is a major source of unfairness on our tests. It is a skill set that is not available to everyone. Some individuals are naturally better at gaming a test, while others pay expensive tutors to learn some of these skills. Unfortunately, the rest of the population must be resigned to the fact that even though they know the content just as well as their fellow examinees, they will likely get a worse score because they are "just not a good test taker." The playing field is far from level.

### Testwiseness has existed since the multiple-choice question was first invented in 1914 and has remained intractable ever since. Why haven't we done more to address it?

The testing industry is generally aware of testwiseness and the unfair impact it has on test scores. Some testing professionals advocate that instead of trying to find ways to eliminate testwiseness, we instead level the playing field by training all test takers equally in testwiseness skills. Unfortunately, widespread and effective training is not yet available, and testwiseness continues to benefit only some.

While we might know about testwiseness, there is a misunderstanding over the extent it impacts test scores. Many testing professionals ignore the issue by telling themselves the influence of testwiseness is negligible. However, research projects conducted over the past few years indicate clearly this testwiseness effect raises scores between 5% to 10%. Sonja Willing and her colleagues at the University of Dusseldorf experimentally compared The Discrete Option Multiple Choice (DOMC)[5] item format with the multiple-choice item format. They showed that when the item content is identical, the test scores clearly show a substantial testwiseness effect.(Papenberg, M., Willing, S., & Musch, J., 2017; Willing, S., Ostapczuk, M., & Musch, J., 2015)

## Testwiseness Cues

- Word association between option and stem.
- Grammatical agreement of option with stem.
- Significant keywords: *Never, Always, Only, None, Impossible* (could be blanket statements).
- General statements juxtaposed with complex alternatives (general may be correct).
- Descriptive statements juxtaposed with vague alternatives (descriptive may be correct).
- Median values in a set of numbers (high and low outliers may be incorrect).
- More than one similar option (both may be incorrect).

**IT IS TIME FOR US TO TAKE A STAND AGAINST ALLOWING SUCH AN IRRELEVANT SKILL HAVE SUCH A SIGNIFICANT IMPACT ON TEST SCORES.**

Over time, this lack of response led to its general acceptance as a necessary evil. It is time for us to take a stand against allowing such an irrelevant skill have such a significant impact on test scores. Luckily for us, the technological innovations offered by DOMC-based SmartItems make effectively tackling testwiseness—and simultaneously addressing other security problems—both painless and easy.

(See Chapters Seven and Eight for descriptions of studies where DOMC is used in research or in actual testing programs, and where it is directly compared to traditional multiple-choice questions.)



*Embracing SmartItems, will dramatically reduce the current expenses that are draining testing program resources.*

## How SmartItems Decrease Costs

Tests today are expensive to create, maintain, and administer. Single test questions can cost as much as $2,000. When these are stolen, they need to be replaced, and the cost compounds. Even in this age of technological efficiencies, test administration costs for computerized tests are higher than they are for paper-and-pencil tests, because we administer them in conditions very similar to the old paper-and-pencil tests:

- Examinees have to travel to another location to take the test.
- Test centers have a dozen or more workstations, rather than the several hundred desks found in auditoriums and gymnasiums used for paper-and-pencil testing. Not only is the computerized testing unscalable, but the workstation equipment costs more than the equipment used for paper-and-pencil tests.
- With computerized tests, the ratio of human effort (mainly proctors) to test takers is higher than it is with paper-and-pencil tests. The cost of proctoring for a dozen simultaneous individuals taking computerized tests is the same for several hundred people taking paper-and-pencil tests.
- Test security costs continue to increase as we strive to keep up with the creative efforts of individuals determined to commit test fraud.

For more than 30 years, technology has been applied to test development, test administration, sending test results, storing test results, scoring, and much more. Unfortunately, technology has not diminished the overall costs of testing—mainly due to our commitment to keeping the logistics of testing the same as it has been for more than 100 years.

**The result is that programs have less money to spend on improving fairness, security, skill measurement, and accessibility.**

Embracing SmartItems will dramatically reduce development and security expenses that are draining testing program resources. Consider the costs associated with writing items. Today, single fixed items are created by SMEs at considerable expense, then written and replaced when the exam is stolen and questions shared. SmartItems can never be compromised, and therefore, there is no need to ever replace them. SmartItems also reduce the costs of test security efforts. By preventing most theft and cheating, programs can save substantial time and money in detecting and responding to security breaches.



*SmartItems will supercharge this trend, making it so all testing programs can deliver exams to any device (computer, smartphone, etc.) at any location (home, school, a public library, a vacation rental, etc.)—all with enhanced security.*

## How SmartItems Improve Convenience

The current process for taking a high-stakes test is inconvenient. From finding a testing location, registering, scheduling, driving to a testing center, checking in, and sometimes waiting weeks for your results, the process is convoluted and time-consuming. This lack of convenience also introduces unfair barriers to individuals who have a difficult time making it to a test center.

With their reliance on the internet and security benefits, SmartItems have the potential to significantly improve convenience for all test takers.

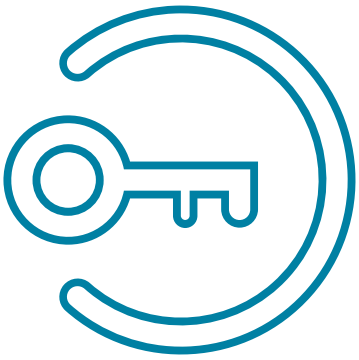
These two features make it so testing programs that use them will be able to administer tests in a broad array of testing locations and at more convenient times. With SmartItems, more and more programs can administer tests in the comfort and safety of an examinee's home—all without testing programs having to worry about exposure and security issues.

SmartItems will supercharge this trend, making it so all testing programs can deliver exams to any device (computer, smartphone, etc.) at any location (home, school, a public library, a vacation rental, etc.)—all with enhanced security.









*SmartItems can increase fairness by making tests more accessible to individuals who require accommodations.*

## How SmartItems Increase Accessibility

As I mentioned earlier in this chapter, establishing validity is the cornerstone of quality testing and trusted scores. Another way to think about validity is through the lens of fairness. When test takers are not given an equal opportunity to demonstrate their abilities, a test is not fair and test validity is damaged. One such source of unfairness is accessibility. All tests should be designed so every test-taker can demonstrate their knowledge and abilities. If a test introduces barriers that make it so some test takers are unable to do this, then that test has validity issues.

Common examples include individuals who have difficulty reading being asked to take a test that uses wordy items; individuals who speak only Spanish taking a test that contains French text; or blind individuals taking a test that presents graphical images.

The concept of Universal Design (UD) was introduced in the 1970s to guide architects to design buildings that are accessible to people of all abilities, cultures, languages, etc. This concept has recently been applied to assessments. We strive to create tests that are accessible to all test takers, regardless of circumstances.

**However, while testing programs today work hard to ensure accessibility, it is a difficult task. SmartItems make it easier.**

The malleable and responsive inherent nature of SmartItems makes them an ideal tool to accommodate UD features. For example, imagine an individual who speaks Japanese as a first language is taking a U.S.-built college entrance test. Typically, such tests are given in English. There are no alternatives for someone whose primary language is Japanese. Such a circumstance benefits fluent English speakers and disadvantages speakers of other languages. However, with SmartItem technology, it is possible to provide the test taker with a choice of languages. A simple change in settings would serve up a more appropriate exam. I'm not suggesting this is easy to do, but it would be tremendously valuable.

SmartItems encourage and support flexibility in item and test designs, both in the content and in how individual test takers can respond. The increased ease of adding accessibility features will eliminate sources of unfairness and provide additional evidence of the usefulness and value of test scores.

# 01 CONCLUSION

The field of testing is plagued by a few critical problems that desperately need to be addressed: test fraud, testwiseness, ballooning costs, inconvenience, and inaccessibility. High-stakes exams are stolen with impunity, and cheating remains rampant; testwiseness still unfairly influences today's test scores as it did 100 years ago; testing is too expensive, which restricts its reach and value; and it remains inconvenient for many to travel to a testing center to take an exam.

We cannot tout the fairness and validity of our exams while these problems still afflict our tests. We owe it to ourselves, our stakeholders, and—most importantly—to the individuals who take our tests, to do everything we can to address these major hurdles. Luckily for us, there is at least one new solution available, if we are willing try it out. These sources of unfairness and others can be addressed by embracing the innovation of SmartItems.

## FOOTNOTES

1. I find it useful to refer to testing as an industry. This includes all high-stakes testing in elementary education, higher education, workforce certification, etc. If it involves the characteristics of creating, administering, evaluating, or theorizing about tests, it is part of this industry.
2. Testing term: A test item is interchangeable for "test question."
3. Testing term: Computerized adaptive tests adapt the test to the ability of a test taker in a way that is similar to how a bar is raised (or lowered) for high jumpers. Those test takers with higher ability are given answer questions of greater difficulty, achieving a score based on the difficulty level they achieve.
4. Testing term: A high-stakes test is any test where the achieved score is important to the test taker or other stakeholders. This could be because the score will result in a promotion, in graduation, or in teachers receiving a pay-raise, etc.
5. Testing Term: DOMC stands for Discrete Option Multiple Choice. It is a type of item where options are presented sequentially, rather than simultaneously, making it impossible to use test-taking skills that rely on comparing answer options.

Chapter

# 02 WHAT MAKES AN ITEM "SMART"?

## THE NINE PROPERTIES OF A SMARTITEM

In the last chapter, we learned that SmartItem technology helps address each of four critical industry needs, plus a few others. We must now turn to a discussion of just what a SmartItem is. What characteristics make an item "smart"? This chapter tackles that question, defining SmartItems as having nine distinct properties.

These properties are best thought of as different ways of understanding what SmartItems are and how they work.

These nine features define and describe an innovation not seen before in the field of testing. It is these properties that allow a SmartItem to address the critical needs described in Chapter One and enable the remarkable benefits discussed throughout this book. Let's look at each of these in turn:

### 9 Properties of a SmartItem

1. A SmartItem is an item.
2. A SmartItem is completely congruent with the breadth and depth of a skill.
3. A single SmartItem can render in millions of ways.
4. SmartItems render using "constrained randomization."
5. Code can be used to create a SmartItem, but it is not the only way.
6. The same SmartItem looks different to each test taker.
7. The purpose of SmartItems is NOT to generate traditional items.
8. SmartItems are not a specific item type. They are a treatment for any item type.
9. SmartItems can be enhanced by the DOMC item format.



## Property #1
## A SmartItem is an item

A SmartItem is an item—a single item—used on exams to measure important, clearly identified skills. Similar to traditional items, a SmartItem is created by subject matter experts (SMEs), is assigned an I.D. number, and is stored in an "item bank" on a computer system. It is evaluated with expert reviews, and eventually with item analyses.

Also like traditional items, a SmartItem can be used with any type of test design, such as linear test forms, computerized adaptive exams, linear-on-the-fly forms, etc. If and when evaluation indicates a SmartItem isn't functioning well, the SmartItem can be revised or retired.

These similarities with traditional items are important. When people first hear about SmartItems, there is a tendency to assume they are something more than items. It is assumed they produce items or are item models. This is not the case. A SmartItem is an item, just an unusually innovative one.

## Property #2
## A SmartItem is completely congruent with the breadth and depth of a skill

Early in my career, when it was my job to create items and tests, the term "congruent" was used frequently. In the context of item development, congruence meant the item being written by a developer was covered somewhere in the text of the skill[1] description.

For example, imagine a test on automobile repair and maintenance requires students to know the skill "Perform a brake system inspection." A traditional test item that is designed to measure that ability could ask specifically about inspecting one type of brake: drum-style brakes. This item would be considered congruent with the required skill, even though it was limited in scope and did not mention any of the other types of brake systems, such as disc-style brakes.

The congruence of SmartItems is different than that of traditional items. In concept, it is closer to the mathematical use of the term "congruence," which includes the term "equal." In geometry, angles of a geometric figure are said to be congruent if they are equal, even if they need to be flipped or moved.

As two equal angles are considered congruent, a SmartItem is an item that is completely congruent with its respective skill. This is an unusual circumstance in the field of testing, and warrants further explanation. Let's look at another example of how traditional items are only congruent with a small part of a skill.

This is a learning objective found on the Internet for a university-level biology course:

> Identify the characteristics and basic needs of varieties of living organisms and ecosystems.

In writing an exam, a course professor would be tasked with creating questions based on this objective. However, a single traditional item would not be able to cover the entire skill. Rather, the professor would need to write a series of questions that touch on the wide range of topics included in the learning objective.

This might include writing one question about the characteristics of a specific living organism, one question about the basic needs of another specific living organism, one question about the characteristics of an ecosystem, and one question about the basic needs of an ecosystem. As you can see, each traditional item written for this objective only measures a small slice of the entire skill outlined in the objective.

In comparison, a SmartItem will measure the entire scope of a skill. The SmartItem for this example item will be designed to measure all four aspects of the learning objective. It will measure the student's ability to:

1) Identify the characteristics of all varieties of living organisms.
2) Identify the characteristics of all varieties of ecosystems.
3) Identify the basic needs of all varieties of living organisms.
4) Identify the basic needs of all varieties of ecosystems.

How can a *single item* do all that?

A SmartItem uses special technology during the development process, including programming code and SME-supplied variables, to enable it to cover the entire objective. However, in addition to using new supportive technology, writing a SmartItem depends even more on the quality of the description of the skill. This includes any constraints or examples deemed relevant.

Let's use another example, one from K-12 education. Here is a Common Core Standard for third grade math (3.OA.A.4).This learning objective states:

> Determine the unknown whole number in a multiplication or division equation relating three whole numbers.

## Learning Objectives

| | |
|---|---|
| Organism Characteristics | Ecosystem Characteristics |
| Organism Basic Needs | Ecosystem Basic Needs |

One SmartItem can render in thousands—even millions—of ways to cover every facet of this complex objective.

### Thousands of Renderings Yield Examples from all Four Learning Objectives

| | |
|---|---|
| Organism Characteristics | Does the Palo Verde tree have green flowers? |
| Ecosystem Characteristics | Select all the characteristics of a desert ecosystem from the list below: |
| Organism Basic Needs | Does a mammal require sunlight to regulate its body temperature? |
| Ecosystem Basic Needs | Which natural resource is most endangered in the desert ecosystem? |

To create an item that covers this entire standard, the item must include both multiplication and division, and must include equations that have three terms, each representing a whole number. Examples given on the website to clarify the standard indicate the whole numbers can be made up of one or two digits. This means the item will only use numbers between 1 and 99 in the three terms of the equation. The examples provided also helped to clarify the three positions of the "unknown whole number" and the two positions for the equal sign (=). The SmartItem, then, must include these variables and their values:

1. Multiplication and division operations, randomly determined.
2. Whole numbers between 1 and 99, randomly determined.
3. Position of the "unknown whole number," three positions being possible in the equation, randomly determined.
4. Position of the "=" sign, two possible positions in the equation, randomly determined.

The SmartItem is then written using code, or the SmartItem graphical interface (see Chapter Six for more information) to include all of these four variables and their values.

Figure 2.0 is a snapshot of what the 17 lines of code look like for this particular SmartItem (I will be discussing this use of code and SmartItem development software in more detail in Property Five.).

When this item is displayed to a test taker, the SmartItem will randomly select values for each of the four variables, and will render differently for each test taker. For this particular SmartItem, thousands of possible renderings exist—thousands of traditional items would need to be written in order to have that same effect and cover the entire skill. To see this third grade math SmartItem in action, go to http://examples.caveon.com/common-core-math.

SMARTITEM IN ACTION:
Click HERE or on Figure 2.0 below.

Each time you click the "reload" button for this SmartItem (located in the top-right corner of the screen), you see a different rendering of a single SmartItem. You should recognize two things right away when interacting with this item:

1. To test takers, this SmartItem looks just like any traditional item. This is because test takers only see a single rendering of the SmartItem each time.

**Figure 2.0**

Stem

```
{% let is_multiplication = choice([0, 1]) %}
{% let idx_to_hide = choice([0, 1, 2]) %}
{% let num0 = randint(1, 9) %}
{% let num1 = randint(1, 9) %}
{% let num2 = num0 * num1 %}
{% if is_multiplication %}
 {% let numbers = [num0, num1, num2] %}
 {% let symbol = '×' %}
{% else %}
 {% let numbers = [num2, num1, num0] %}
 {% let symbol = '÷' %}
{% endif %}
{% let hidden_number = numbers[idx_to_hide] %}
{% let numbers_for_stem = numbers.copy() %}
{% do numbers_for_stem.insert(idx_to_hide, '?') %}
{% do numbers_for_stem.pop(idx_to_hide + 1) %}
{{ numbers_for_stem[0] }} {{ symbol }} {{ numbers_fo
```

**Common Core Math**

**Objective: Determine the unknown whole number in a multiplication or division equation relating three whole numbers.**

**35 ÷ 5 = ?**

Choose 1

- 6
- 8
- 9
- 7
- 5

**2** After reloading the item several times, you will notice the SmartItem renderings do in fact cover all four variables outlined in the learning objective. The result is a single item that covers the entire specified skill. Compare this to a traditional item, where the item writer would have to decide on which specific values to write items for, and each would only measure a fraction of the specified skill.

Whether the SmartItem is considered at the micro or macro level, the concept of congruence is clear: **A SmartItem must completely cover stated aspects of a skill description. There can be no gaps.** The importance of this property cannot be overstated. It allows teachers, parents, students, administrators, and others to trust that individuals will be tested on the skills that have been agreed upon by all stakeholders.

## Property #3

## A single SmartItem can render in millions of ways

As mentioned in the previous section, a SmartItem displays differently to every test taker. Each unique way a SmartItem is displayed is what we call a "rendering." Depending on the number of variables and other factors, a SmartItem can render in thousands, hundreds of thousands, or many millions of ways.

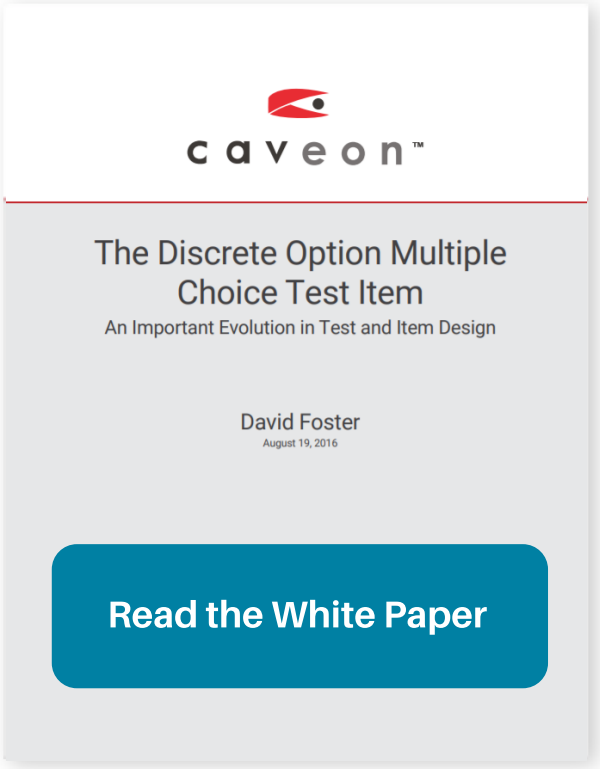
Here is an example. Consider this learning objective adapted from an online university's American History course:

> Identify amendments in the United States Constitution and Bill of Rights that protect civil liberties and civil rights.

There are 27 amendments to the U.S. Constitution, 17 of which were created to protect civil liberties and civil rights. The SmartItem for this example must include the 17 amendments and the civil liberties they protect. (Example: The Thirteenth Amendment outlaws slavery.)

This SmartItem could be written using a preponderance of options and the DOMC item type[3]. (To learn more about the DOMC item type, please see the below white paper, or go to http://bit.ly/DOMCwp.) The DOMC-enhanced SmartItem is described and demonstrated below. To start, the item writer (in this case, we'll assume that is me—but it could be done by anyone) will simply write the stem:

> "Is this statement true about how the amendments in the Constitution protect civil liberties and civil rights?"



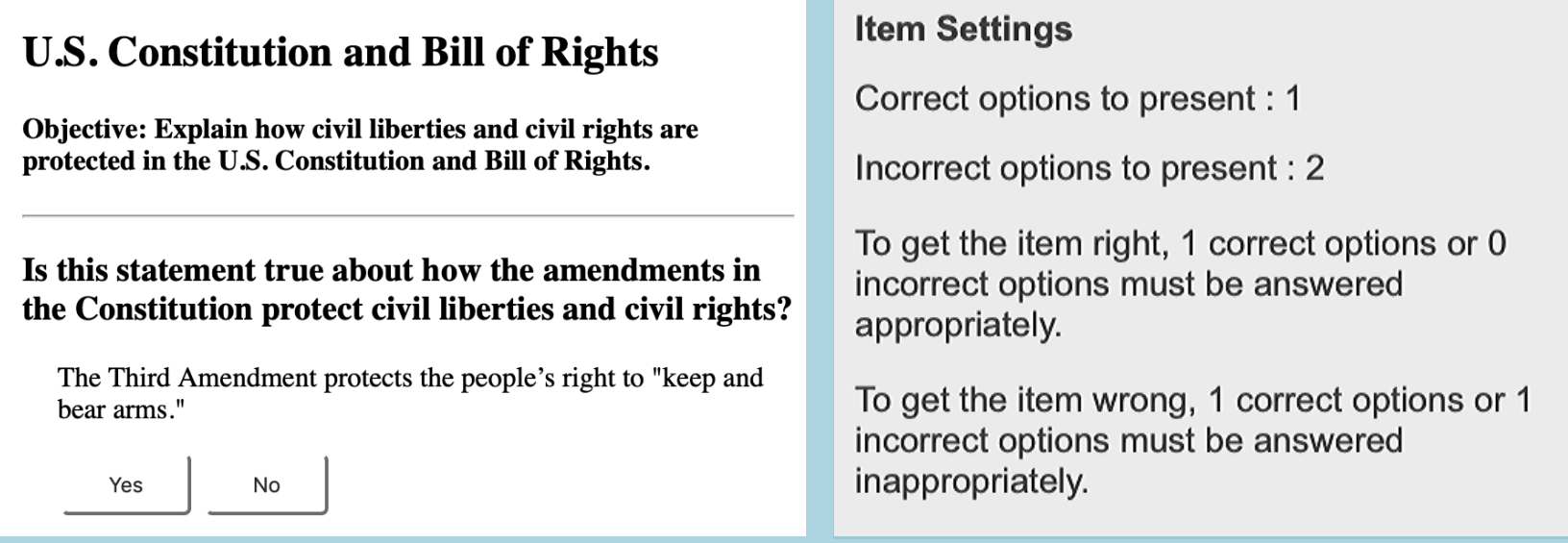



**Figure 2.1**

I will then include all 17 correct amendments as correct options, and create as many incorrect options as I would like by combining the amendment number with an incorrect protection statement. (Example: The Twelfth Amendment outlaws slavery.) You can read more details about this SmartItem in Chapter Five.

SMARTITEM IN ACTION:
Click HERE or on Figure 2.1 .

In the end, the finished SmartItem must include these variables and their values:

- The 17 correct and nine incorrect options.
- The display settings that select one correct option and two incorrect options randomly from the pool of 26 options.
- The DOMC format that may present only a subset of the three options.
- The different randomly-ordered presentation of the options.
- The inclusion of a Random Additional Option (RAO[2]; peculiar to the DOMC format), unscored, presented with a probability of .5 when the item is scored.

Given these conditions, the example of this SmartItem can render in a whopping 413,406 ways! You can view an example of this SmartItem by clicking above, or going to http://examples.caveon.com/us-constitution.

The SmartItem's abilities to render in so many different ways and to cover an entire standard or skill enable it to effectively address item theft and cheating.Test security threats related to item harvesting are neutralized if a test taker is unable to predict what they will see on the test. Not only does a SmartItem cover all the content of a skill, but the sheer number of renderings it's capable of producing make it impossible for any test taker to share or gain useful pre-knowledge about test content.

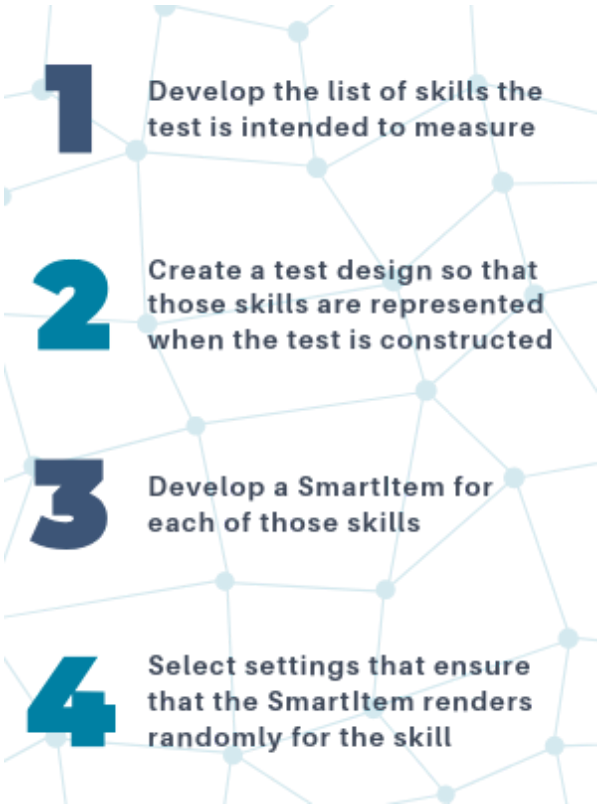
The 413,406 different possible renderings of the above item on U.S. Amendments make it extremely unlikely that two test takers will see the same content. This makes sharing questions or posting them online a pointless effort. The shared information will simply not be useful to the next person taking the test. With so many renderings, you can no longer cheat by sharing test content, by buying questions and answers, by asking a friend to take the test and tell you what they saw, etc.

The only way to prepare for a test that utilizes SmartItems is to improve your knowledge of the skills being tested. The examinee must properly prepare by setting aside the necessary time to learn about and understand the amendments[4]. That's the only way to score well on a SmartItem-based exam.

## Property #4

### SmartItems render using randomization stratified by skills

Each rendering of a SmartItem is created and presented randomly. While it is constrained to the limits of the skill being measured, the SmartItem's form of randomization ensures the entire scope of each skill can be represented on the exam. This approach is not new. Often referred to as "stratified randomization," this approach is used by scientists conducting experiments across a wide array of fields. For SmartItems and test design, there are four steps to randomization:




The randomization of a SmartItem is constrained within each skill. If a particular skill is given greater weight, the test designer can arrange for the SmartItem related to that skill to be administered more than once on the test.

Randomization, stratified or not, has another interesting and positive effect. Test takers have no choice but to prepare across the entire breadth and depth of a described skill, as the specific content measured within that skill will be unpredictable. It is not just the test takers who will feel this pressure to better prepare across the entire set of skills, but also those who teach or create training materials.

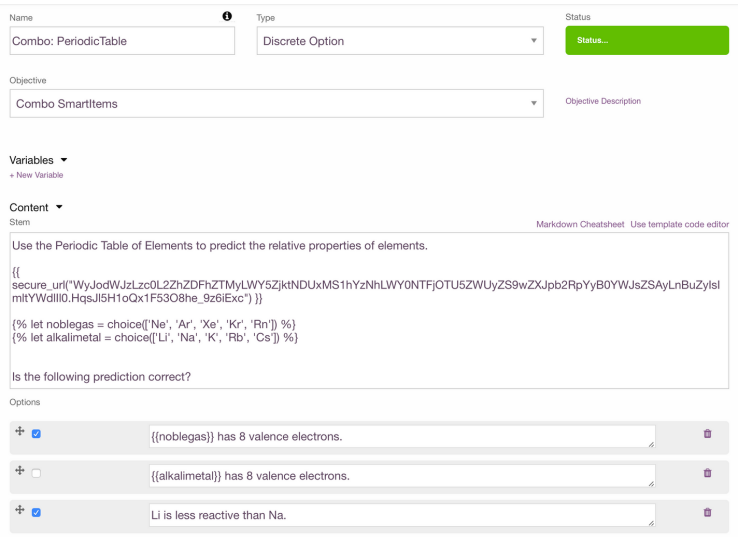
For example, all K-12 teachers will give up trying to teach their students only the skills that are most likely to be seen on state assessments. Rather, they will begin to "teach to the test" by ensuring their students thoroughly know about and understand the intricacies of the skill that could be referenced on the test. Instead of the pejorative connotation that the phrase implies, "teaching to the test" will be used to indicate a higher level of teaching expertise. This effect is not limited to K-12 educational tests, but will occur wherever high-stakes tests are administered.

## Property #5

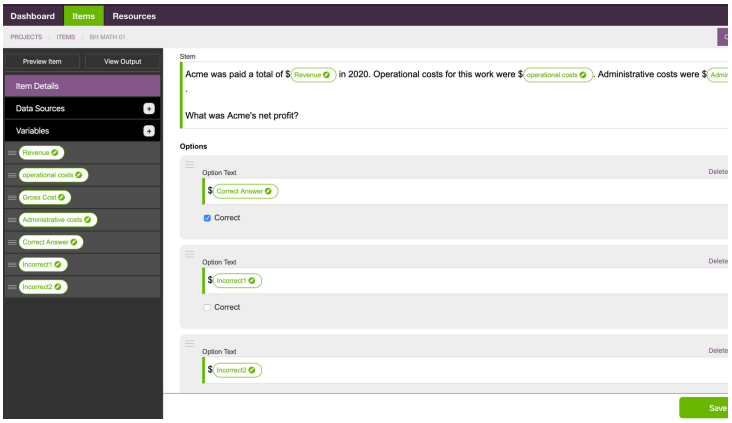
### Code can be used to create a SmartItem, but it is not the only way

As we learned earlier in this chapter, the goal of SmartItem development is to cover the skill completely. This is accomplished by using technology. Each of the SmartItem examples mentioned in this book have been built by using code, by utilizing the SmartItem Graphical User Interface (GUI), or by creating a preponderance of options.

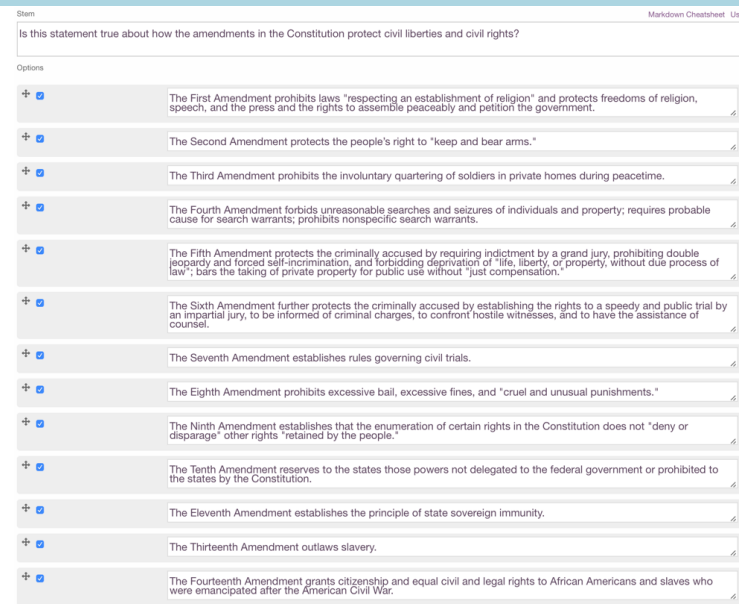
**Figure 2.2**




Option 1: Using code in stem or options




Option 2: Using a graphical user interface




Option 3: Using a Preponderance of Options

## Option #1 | Using Code in Stem or Options

Code is used to program an item so that it varies in the ways necessary to cover the identified skill. Imagine a skill asks you to identify the U. S. Presidents before, during, and after the U. S. Civil War. In order to cover the objective, the SmartItem would need a variable to cover those three time periods, as well as variables that include every president within those periods. Code is written for each of these variables, then an item format is selected, and the SmartItem can be run, producing a number of renderings. Running a SmartItem is the same as running any coded program. When the SmartItem is presented to a test taker during an exam, the code will be executed.

## Option #2 | Using a Graphical User Interface

It needs to be noted that coding skills are not necessary for creating this type of SmartItem. There are SmartItem tools available that make building this type of SmartItem easy for every item writer SME. This GUI software lets individuals with no coding experience create SmartItems using a powerful embeddable editor. This editor enables the SME to design variables and values, and to place them into the item components using a drag-and-drop interface. Without writing a single line of code, the SME can create a SmartItem that functions as if it was coded by hand.

## Option #3 | Using a Preponderance of Options

The other method for creating a SmartItem is to write a sufficiently large number of options. Because options are required for this non-code alternative, it can only be used with item formats that are of the selected-response type. These include true/false, DOMC, multiple-choice (MC), and all their varieties. This method for creating SmartItems is especially appropriate when measuring the ability to recall information about a domain of knowledge that is relatively unambiguous. To create a SmartItem using this method, you simply write as many options as needed to make sure the SmartItem covers the required skill. For some questions, there might be 20 options. For others, there could be hundreds or thousands.

The decision on which method to use should never hinge on whether an SME knows how to code. One method is not inherently better than another. Rather, some skills are better suited to a coded question, some to a GUI interface, and others to a question with many options. Also, depending on the skill, it can be easier to use one method compared to the other.

It's important to realize that these methods are not mutually exclusive, but that they can be combined into the same SmartItem, as well as across an entire test. In addition, please note that SmartItem design and development is very new, and I'm sure that new advances for creating these types of items will be developed over time.

Let's take a moment to view another example of a non-coded SmartItem that uses a preponderance of options rather than code. Consider this learning objective from a chapter on psychology research methods found in an introductory psychology textbook.

Read and be able to recall the major points of the Research Methods chapter

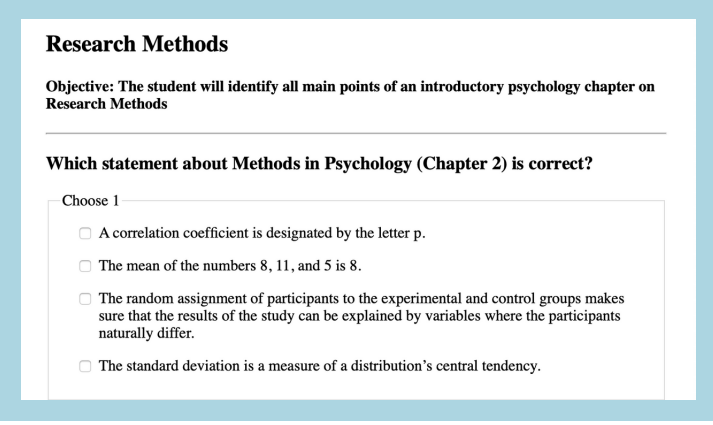



**Figure 2.3**

SMARTITEM IN ACTION:
Click HERE or on Figure 2.3.

I created this SmartItem in about six hours (you can view it in action by clicking the link above, or by going to examples.caveon.com/research-methods). As a decent SME on this topic, I read over the chapter carefully and selected the significant points included by the author. For each of those points, I wrote one or more correct statements and one or more incorrect statements. The SmartItem ended up with 67 correct options and 73 incorrect options. This is a DOMC item, so for each test taker, one correct option and three incorrect options are randomly selected to potentially be presented during a test. Due to the design of DOMC, perhaps not all of the four options will be presented. The DOMC item may end early when the rendering is answered correctly or incorrectly.

By reviewing this example, you can see how a DOMC (or MC) item with many options can be used to create a great SmartItem. Of course, remember that not all questions with a lot of options—and not all DOMC-based items—are SmartItems. To qualify as a SmartItem, the item must cover the skill completely.

## Property #6
## The same SmartItem looks different to each test taker

Imagine two test takers are sitting side-by-side and are administered the same SmartItem at the same time. While these test takers have both been assigned the same SmartItem, they will each see a different rendering of that SmartItem. The item is the same, but the specific content in each rendering is different.

As has been stated in previous sections, certain variables of a SmartItem are intended to change randomly for each test taker. However, even though the content of a SmartItem is variable, the format of each rendering is recognizable and familiar for individual test takers. For example, if the SmartItem uses the MC format, then all test takers will see an MC item. Test takers are generally not aware of, nor concerned by, what is happening behind the scenes, or by what other test takers might be experiencing.

Figure 2.3 shows us different renderings of the same SmartItem. Notice how these renderings look like a normal MC item, how the format stays the same, and how certain content changes. Behind the scenes, there are variables being manipulated. These are: (1) the short story, (2) the words that are underlined, (3) the question that is asked. To be congruent with the skill description or objective, the question may also display renderings that ask about punctuation and vocabulary.

In addition to the values of particular variables that change with each rendering of a SmartItem, there are certain characteristics that also change. These characteristics include:

- Difficulty- Some renderings will be more difficult than others, but will remain within the limits of the assessment objective.
- Discrimination- Some renderings will discriminate better than others.
- Guessing- Depending on the item format used, some renderings may be easier to guess correctly.
- Length of Content- Some renderings will take more time to read, consider, and answer than others.

**Figure 2.4**

**EXAMPLE #1**

Read the paragraph and complete the task that follows.

Variable — We went to visit my uncle in Massachusetts. He lives just outside of Boston, so we went sightseeing in the city. We went to see were Paul Revere lived. He was a famous hero in the American Revolution. After that, we walked through Boston Commons, a large park. Then we went to a baseball game.

Variable — Choose the word that is incorrectly spelled.

- uncle
- through
- sightseeing
- were

**EXAMPLE #2**

Read the paragraph and complete the task that follows.

We went to visit my uncle in Massachusetts. He lives just outside of Boston, so we went sightseeing in the city. We went to see where Paul Revere lived. He was a famous hero in the American Revolution. After that, we walked through Boston Commons, a large Park. Then we went to a baseball game.

Variable — Choose the word that is incorrectly capitalized.

- city
- Revolution

Variable

- Park
- uncle

- Reading Level- The text of some renderings will be of a higher reading level than others.

Because these characteristics are mostly dictated by the content of the skill, the overall range of these differences will likely be small, but they will still be significant. They will likely still impact test taker performance, and therefore, the test taker's score.



We must take a moment to address a concern I am sure you are feeling. How can these tests be fair if test takers do not see the same content? And if that content varies in many ways, including difficulty, how is that justified? It is naturally alarming to think one test taker will receive a more difficult rendering of a SmartItem than another. And that observation is correct; not only might the difficulty be different, but other item characteristics—such as discrimination, amount of text, reading level, complexity, average time to process the item, language, and ability to use test-taking cues—might also vary.

At this point, I ask you to put aside your concern about these differences. I will address them soon. I know it's hard. For more than a century, identical items and tests have been at the core of the standardization process. However, this has been changing. Over the past few decades, there have been plenty of examples where the requirement of strict standardization has been modified.

Chapters Four, Five, and Six are dedicated to explaining these differences. More than that, they will show that when truly random, these differences become a part of the solution rather than the problem. When used correctly, differences contribute to better measurement.

Property #7

## The purpose of SmartItems is not to generate traditional items

One of the most common questions I receive about SmartItems has to do with Automated Item Generation (AIG): "Aren't SmartItems just another form of AIG?" The answer is a definite "No." SmartItems are built to be used on tests.

Creating SmartItems is not equivalent to the use of AIG. AIG is a 20-year-old movement designed to create a large number of traditional items. With AIG, technology is used to inexpensively build a pool of items that can be used to (1) expand pools for adaptive testing and other test designs, and (2) replace compromised items and tests. AIG users and researchers have the basic hope that items created following particular procedures can be used without expensive reviews and field testing. Predicting item parameters, and therefore predicting item performance in exams without other information, is a goal of many AIG users. There are many AIG approaches, and several textbooks and reference books that describe those approaches.

SmartItems can be used to automatically generate other items, or to replace compromised items[5]. However, the primary purpose of SmartItems is to use them, on a test, to render on-the-fly to test takers.[6] The value of SmartItems in this context is to serve as a preventative security method, not a reactive one. SmartItems make tests un-compromisable. When built according to the principles outlined in this book, SmartItems make item harvesting a useless endeavor—and have that same effect on many other forms of cheating as well.

## How are SmartItems Different from AIG?

1. The purpose of AIG is to create a large number of items using computer technology. The purpose of SmartItems is to create only a single item per important skill.

2. AIG uses item models patterned after strong items or theory. These models are not ever used on a test. SmartItems are the items that are used on a test.

3. Items built from AIG are often not immediately useful. They may require reviews by language and psychometric experts, and may need to undergo field testing. SmartItems can benefit greatly from such field testing, but each rendering does not require the same level of scrutiny.

4. Items created by AIG must have predictable parameters. Many AIG researchers still prefer that the replaceability of items requires identical statistical parameters. SmartItems, on the other hand, celebrate and depend on the natural variability of the SmartItem renderings for the safe, competent, and useful measurement of skills.

5. AIG is used to expand the size of an item pool, while SmartItems are intended to contract the size of an item pool.

6. AIG, so far, is used to measure relatively simple skills and perhaps because of that only uses the simplest item formats, generally the multiple-choice format. SmartItems are format agnostic, working with all item types and using whichever format is considered best to measure the described skill.

7. The security focus of AIG items is to provide a large pool of items to either forestall exposure problems, or to replace items that have been compromised or over-exposed. The security focus of SmartItems is to prevent theft and cheating at the outset. Exposure of the SmartItems themselves is irrelevant for security purposes.

8. AIG is often referred to as “mass production” because of its similarity to manufactured goods being bulk-produced. The creation of each SmartItem is a creative, singular effort that is meant to produce a custom product. SmartItem creation is much more analogous to making one-off products such as a work of art, a music composition, or the training and development of a gifted athlete.

## How are SmartItems Different from AIG?

**9** The singular goal of AIG is to create a large inventory of items, hopefully reducing the per-item cost of development. In comparison, SmartItems have several goals:

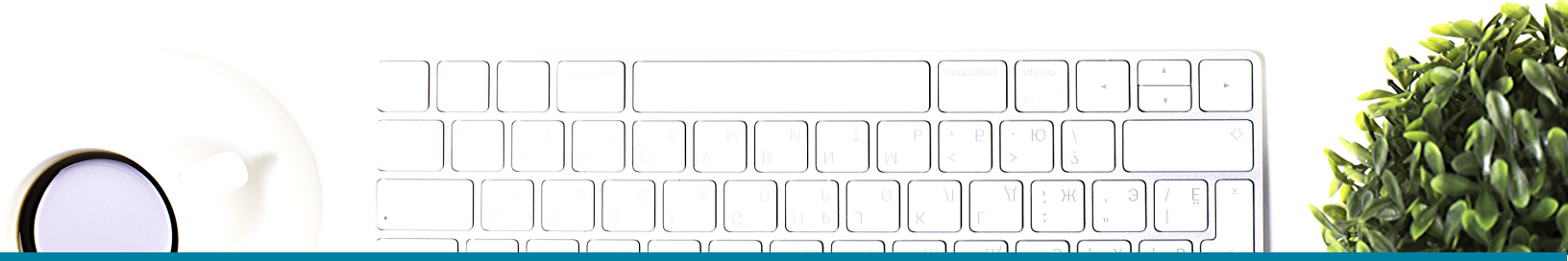
- Preventing cheating and theft.
- Reducing costs associated with security breaches.
- Eliminating expenses related to over-exposure (SmartItems can't be compromised, and as a result, there are almost no maintenance costs after the SmartItem is created. It only needs to be updated if and when the skills it covers are changed or revised).
- Motivating the ideal form of test preparation and instruction (through learning and experience in the skills the test covers).

**10** Both AIG and SmartItems intend to reduce costs. However, they accomplish this goal in different ways. AIG reduces the cost of item development by having computers perform the task that is traditionally done by SME's; SmartItems reduce costs by only needing to be created once, and don't need revising unless the language describing the skill is changed.

**11** AIG supports test design practices that use CATs or test forms. SmartItems use only one form. SmartItems differ in difficulty and can be used in CAT pools.

**12** For AIG, SMEs are not asked to create items, although they may be needed to help design the item models. For SmartItems, SMEs are critical as they remain the only individuals with the experience to know what the skills are referring to, and how to best address them. SMEs are expected to use intelligent tools to create the SmartItems, or work side-by-side with coders who can implement the SMEs designs.

**13** AIG uses item models crafted from existing high-performing items, cognitive science theory, and psychometric theory. SmartItems are created to directly apply to a stated measurement goal.



Property #8

## SmartItems are not a specific item type, but a treatment for any item type

SmartItems are not another item type or item format. SmartItems are not equivalent conceptually to multiple-choice, DOMC, or other selected-response or constructed-response formats. Instead, SmartItems can be applied to any of these. In fact, SmartItem technology must be applied to an existing item format to work at all.

The examples you have seen so far in this chapter have shown SmartItems working as short answer, MC, and DOMC item types. However, SmartItems work for all other item types as well, including the Build List format that Caveon's tools support.

For this item type, seen in Figure 2.4, the test taker needs to cull a long list by dragging the list's line items and dropping them into a different box. Then, the test taker re-orders the second list by dragging and dropping the line item into a new correct order. The SmartItem version of this item type randomly presents different scenarios, each with unique line items. The test taker is tasked with dragging and dropping each variable into a new list in the correct order.

**BUILD-LIST SMARTITEMS**

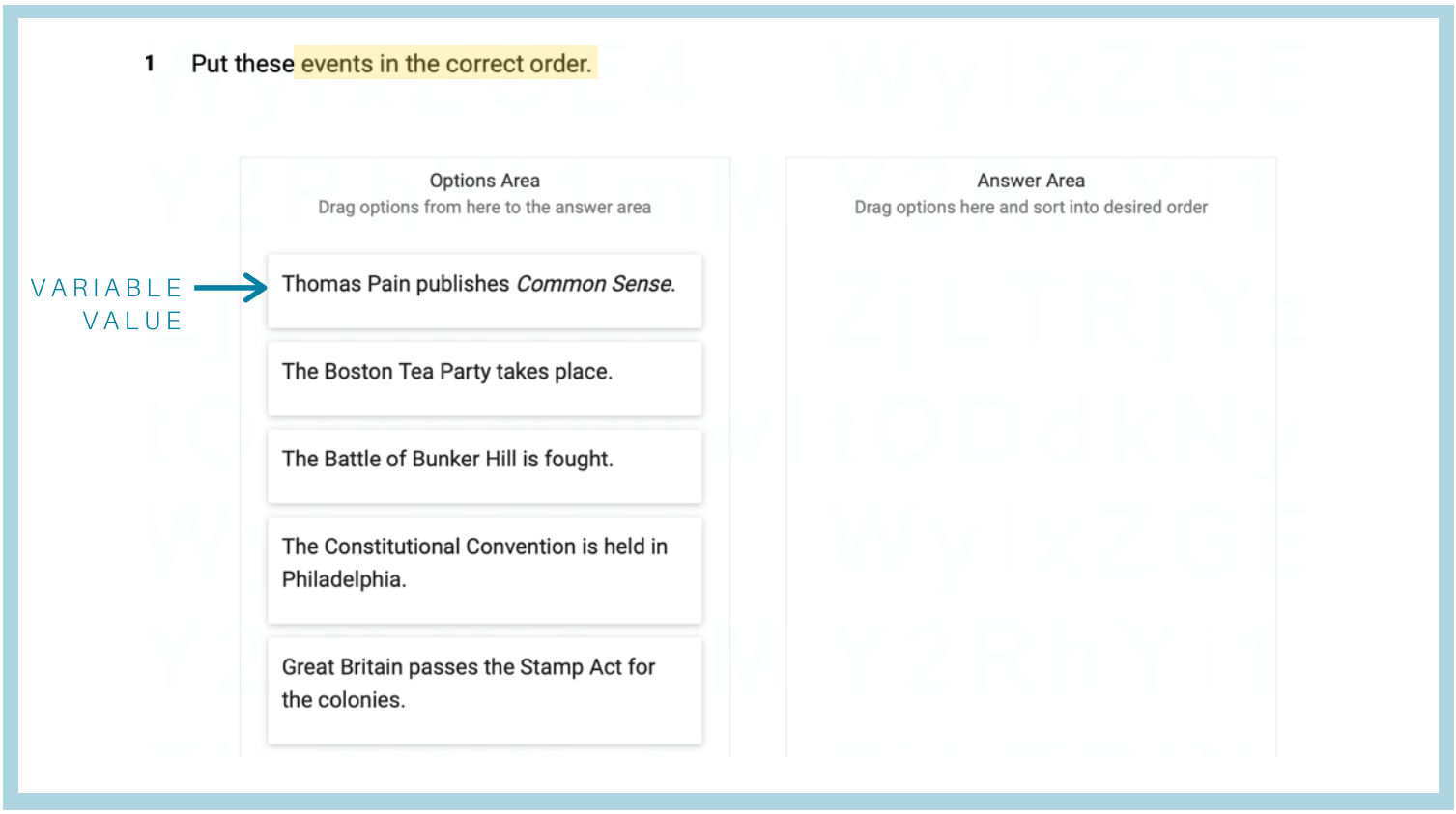



**Figure 2.5**
A Build List item where the alternatives on the left are moved to the right and then re-ordered appropriately, in this case, from earliest to the latest.

**BUILD-LIST SMARTITEMS CONT.**

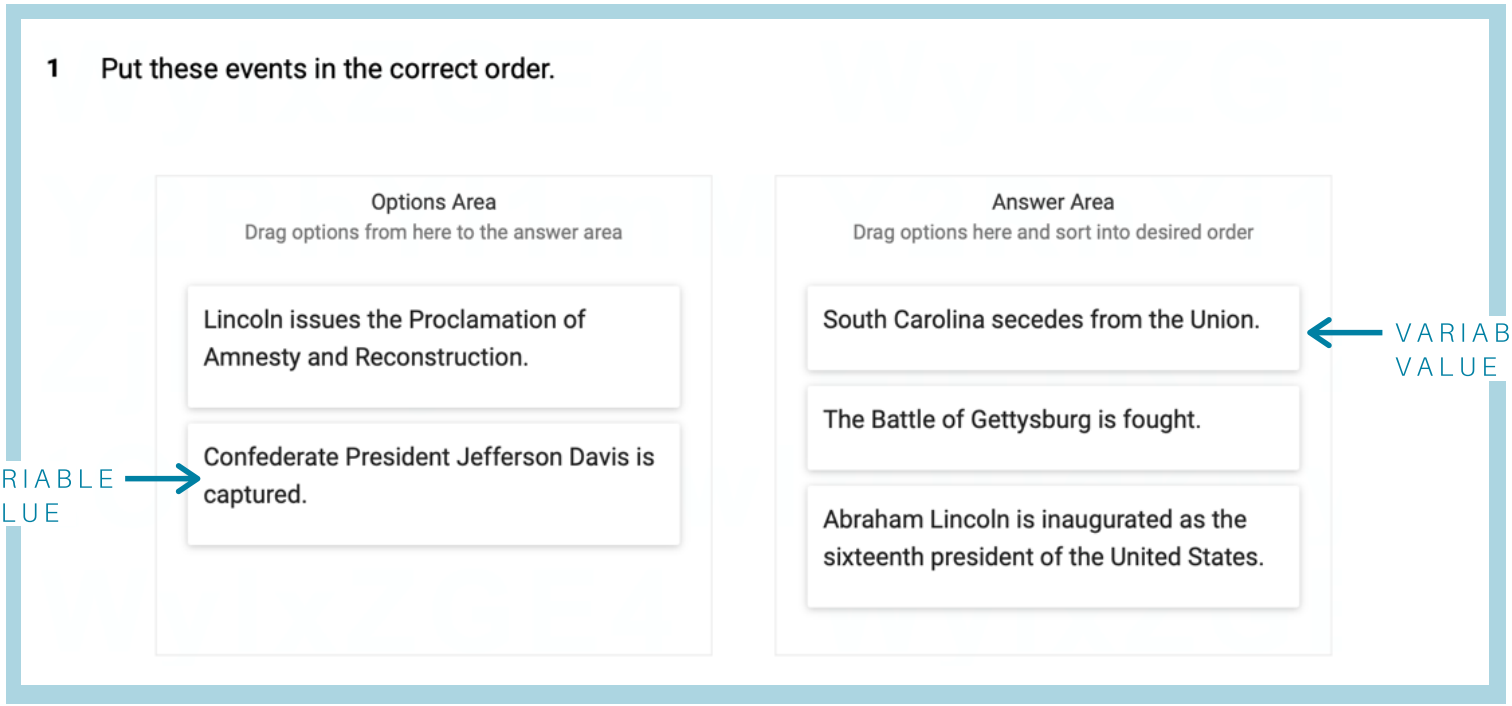



**Figure 2.5** Continued
A Build List item where the alternatives on the left are moved to the right and then re-ordered from earliest to latest.

Figure 2.5 shows a different rendering of the SmartItem, but this time with some work done by the test taker, as some of the choices are moved and re-ordered.

As with any other type of item, SmartItems can make use of text, audio, video, animation, and simulations files. The most important factor is that the item type and any included media are selected because they are best suited to measure the prescribed skill.

## » Property #9
## SmartItems can be enhanced by the DOMC item format

A MC-based SmartItem can be supercharged by using the DOMC item format. While SmartItems can prevent theft and many types of cheating by covering the entire skill and having lots of renderings, it is only by employing the DOMC item format that you can effectively eliminate testwiseness as well. As was discussed in Chapter One, the multiple-choice item type enables and encourages the use of testwiseness skills, helping test takers who are good at using test-taking cues get a better score. Even for high-quality exams, the testwiseness effect can inflate an individual's test score between 5%-10%.

On the other hand, DOMC enables a multiple-choice-based SmartItem to eliminate almost any unfair advantage a test taker may gain because they are test-wise. DOMC allows you to have your cake and eat it too, removing the two largest sources of systematic bias in test scores: cheating and testwiseness. By controlling these two major bias effects, a test score becomes a more meaningful representation of an individual's knowledge of the skills being measured by the test.

# 02 CONCLUSION

These nine properties are what make a SmartItem so very "smart." As you have learned, SmartItems are unlike any other item used on any type of test anywhere in the world. SmartItems are a unique concept, born out of need, and made possible by new technology. The chapters that follow will build on this foundation and provide logical, theoretical, and empirical support for the use of SmartItems in every testing program where security and quality measurement are a priority.

## FOOTNOTES:

1. The term "skill" can also be referred to by near-synonyms such as: ability, performance, competency, objective, proficiency, construct, standard, cognitive process, etc.
2. DOMC Term. A RAO is an extra option that is presented at a set probability when an item is completed and scored. Its use is to make the actual completion of the item opaque.
3. The Discrete Option Multiple Choice item format, or DOMC, is a selected-response format first introduced in 2009 (Foster & Miller, 2009). It is similar in most respects to the multiple-choice format, except that the options appear sequentially rather than all at the same time. Test takers answer YES or NO as to whether they think the option is correct or not.
4. Proper preparation refers to reading textbooks, attending classes, completing assignments, gaining experience, etc. A SmartItem-based test motivates teachers to teach the skills completely.
5. SmartItems are being used by some programs to create thousands of other items, in the spirit of AIG. The renderings of a SmartItem are captured, and treated as individual fixed items. These resultant items are given ID's and stored in an item bank to be used at a later time.
6. The list is based on my understanding of AIG after reviewing all books and articles on the topic available at the time I wrote this. It is possible others in the industry have other opinions about AIG and what it is that would conflict with this view.

Chapter

# 03 THE DIFFICULTY QUESTION

## LOGICAL, THEORETICAL, PRACTICAL, AND EMPIRICAL ARGUMENTS FOR SMARTITEMS

People have many questions about SmartItems, but the most common (by far) has to do with difficulty. A SmartItem renders in a large number of ways, and most of those renderings differ in difficulty. How can SmartItems be fair when two test takers might randomly see items of varying difficulty?[1]

In addition to questions about the differences between individual SmartItem renderings, there are concerns about general difficulty. Could one test taker's overall exam end up being more difficult than another's?

The next few chapters will attempt to answer these questions.

What must first be understood is that the concept of SmartItems is shocking to many testing professionals. The initial reaction of a person when they first learn about SmartItems is one of surprise. How can an item render randomly to each test taker? This kind of item has no precedent. It has never been seen before on exams (especially high-stakes exams) where questions have always been static, fixed, immutable, and exams are always identical. After a moment of contemplation, these professionals usually wonder aloud whether the variability in renderings might have an unfair result. They wonder whether the effect of randomness is tracked in such a way that final test scores can be statistically adjusted to be comparable.

The commitment to fairness rightfully belongs at the center of all measurement efforts in the field of testing and assessment.

Certainly, tests must be fair to test takers so the scores can be compared to each other or a fixed pass/fail standard. This principle is clearly set forth by the testing *Standards*[2] all professionals in this field follow. However, for decades, testing experts have tended to focus on item difficulty as the most important characteristic for evaluating new concepts that

might impact exam fairness and validity. Yet item difficulty is not the only characteristic an item possesses, nor is it the most important. Focusing on this singular feature of difficulty is insufficient to understand the overall impact of an item on a test score.

While the difficulty of individual SmartItems may vary on a test, it is the design of SmartItems—as well as the presence of a significant number of SmartItems on an exam—that makes them fair. To help illustrate this point, this chapter will address this "difficulty question" from several perspectives: logical, theoretical, practical, and empirical. These arguments will help us understand how, with sufficient numbers of SmartItems on a test, the randomness of individual SmartItem renderings creates a balance of difficulty. In doing so, it doesn't differentially affect test scores in a meaningful or unfair way.

## SmartItems | Logical Arguments

### Logical Argument #1

The *Standards for Psychological and Educational Testing* does not, in fact, require that items be equally difficult for all test takers.[3] It is common practice in all areas of high-stakes testing to present different forms of a test to different test takers. Each of these forms—whether they are common fixed-length forms, forms constructed on-the-fly (such as a CAT), or forms produced by some other method—present items of different difficulty for the same target skill. It is nearly impossible to create multiple forms where the items for the same target skill area are equivalent in difficulty.

This general approach of using items with different difficulties on tests is not only acceptable, but it is considered best practice. SmartItems should be considered in this same light.

READ THE "FAIRNESS" ISSUE OF THE LOCKBOX, THE TESTING INDUSTRY'S DIGITAL SECURITY RESOURCE

READ NOW



## AUTHOR'S NOTE:
## **THE ISSUE OF RANDOMIZATION.**

Randomization is prevalent throughout the entire test development and administration process. Here are a few examples:

- Randomization is typically used in the construction of good test forms[4]—forms either equivalent by design, or that can be statistically equated after the test has ended.
- Randomization is used during development to assign items to forms, avoiding the effects of selection bias by test developers.
- Randomization is part of algorithms for the administration of CATs. These algorithms select items from an available pool to present to test takers. These algorithms are an important part of efforts to balance the exposure rates of items.

The randomization utilized in these common test designs is referred to as "stratified random sampling." This is a method of stratifying on one variable first (such as a person's ability or a content area) and then randomly selecting an item to be presented from a corresponding pool.

In all cases where randomization is used as part of the construction of a form, whether during development or test administration, the items given to test takers from the same content strata will differ in difficulty. Despite the expected variance in item difficulty, the forms are considered equally difficult. If there are minor differences in form difficulty, this can be accommodated. Measurement professionals already accept randomization as a reasonable and necessary part of form creation.

It is evident the difficulty question we are addressing throughout this chapter does not apply to only SmartItems. It can be asked about items in virtually all published high-stakes exams. Items always vary in difficulty across forms, because no two forms have exactly equivalent items. Nevertheless, forms can be equated before—or even after—test administration to make sure test scores are comparable and fair. This same process applies equally to tests created of SmartItems.




## Logical Argument #2

Item difficulty variation is an inseparable part of every test taker's experience—even when test takers view identical items on the same test form. This is based on the concept of "perception" in psychology. Here is a paragraph from an online Introduction to Psychology course by Lumen Learning:[2]

> *While our sensory receptors are constantly collecting information from the environment, it is ultimately how we interpret that information that affects how we interact with the world. Perception refers to the way sensory information is organized, interpreted, and consciously experienced. Perception involves both bottom-up and top-down processing. Bottom-up processing refers to the fact that perceptions are built from sensory input. On the other hand, how we interpret those sensations is influenced by our available knowledge, our experiences, and our thoughts. This is called top-down processing.*

While test questions may (though not necessarily) be "sensed" in the same ways by test takers, an individual test taker's different knowledge, and other factors, will modify the processing of the sensory input from the item. Sensory input would comprise the words and images test takers see, or the sounds from an audio file they hear. Examples of other factors that might affect the sensory processing of an item include the temperature of the room, the comfort level of the desk and chair, or unusual sounds from outside of the testing room.

This is a vitally important point to understand. It counters the prevailing assumptions of testing professionals that an item—so long as it doesn't vary in content—is perceived and treated identically by test takers. The basic principles of psychology undermine this assumption. **Rather, individual perception will impact item difficulty, even if the item has fixed content and if the test takers are equally capable.** Their perception of an item will be affected by many factors, including the test taker's history of taking tests, culture, maturity, illness, fatigue, motivation, reading ability, test-taking ability, anxiety, primary language, familiarity with the testing modality, and many, many others. This more accurate conclusion is supported by decades of research in cognitive science.

## SmartItems | Theoretical Arguments

"Can test theory support the use of SmartItems?" After a careful and extensive review of test theory literature, the answer to this question is a conclusive "Yes." As described, SmartItems render randomly within the constraints of each skill being measured by the test. This is the single most important property of SmartItems, and it is supported by both classical and modern test theory. Let's not forget that the two previous arguments showed that tests work well when randomization is used to create forms (on-the-fly or pre-made). In these instances, items differ in many ways, including difficulty. Test theory supports those existing testing models, even when item difficulty is not controlled or accounted for. Why wouldn't that be the case for SmartItems as well?

SmartItems are a new and potentially disruptive technology. Because of this, it is especially important to note straight from the beginning that test theory, both current and traditional, supports the use of SmartItems. The practical use of a SmartItem depends upon the acceptability of randomization—and only randomization—in the different renderings of views of each SmartItem to test takers. You need to remember that each SmartItem is congruent with its skill description, so SmartItems are not randomized across an entire test. Rather, they are randomly stratified by the specific skills being measured. (For a more complete treatment of randomization, please see Chapter Four.) SmartItems use randomization to present unpredictable—but content congruent—renderings of item content to each test taker. It does this for each SmartItem on the test, and therefore, for each skill measured by the test.

Randomization, as used in SmartItems, is completely unbiased with regard to test takers. This is the key point. It means SmartItems will not consistently render more difficult versions for one test taker and provide easier renderings to another. Imagine you and I are two test takers, and we are each taking a 100-SmartItem test. Let's also assume this test is comprised of the same SmartItem duplicated 100 times[5]. The test begins, and we each see the first rendering. Initially, you see one of the more difficult SmartItem renderings (one that is more difficult because of its content, its readability, etc.).

At the same time, I see one of the easier renderings. We both answer and then await the second item. The second question for each of us is different in difficulty as well, but now it is more difficult for me than for you. Variations in difficulty (along with the many other characteristics of a SmartItem's rendering) begin the inevitable path to balance. This is the critical point—at some point over the course of the 100-SmartItem test, the difficulty of SmartItems will balance out. Inevitably, our tests will be of equivalent difficulty, and our test scores will be comparable. (Chapter Nine presents simulations to demonstrate this phenomenon.)

The key theoretical component for this explanation is the error term in the well-known Observed Score equation:

Formula 1:

Observed Test Score = True Score + Error

Any measurement effort, whether in our field or any other science, is inexact. The mathematical relation accepted by most testing measurement professionals illustrates that an Observed Score (OS) from a test is the combination of two factors: a person's True Score (TS) and Error (E).

In this equation, the OS is the observed test score for a single individual. The TS represents the true ability of the test taker concerning the target skills the test is designed to measure. The E represents error, the influence of factors which are independent of the TS. These factors have an impact on the OS, altering it to be higher or lower than the TS.

However, some testing professionals find Formula 1 to be too simplistic, and have thus expanded it to include more than one type of error. In this more refined formula, E is divided into two types of error: Systematic Error (SE) and Random Error (RE).

The way that measurement error is considered in our field, as well as in all other sciences, the equation is better re-written as:

**Formula 2:**

Observed Test Score = TS + (SE + RE)

To understand this formula (and as a result, to understand the characteristics of observed test scores), it is critical to understand the difference between the two types of error and their vastly different effects.

**QUICK REFERENCE GUIDE**

**OS** = Observed Score (TS + E)

**TS** = True Score

**E** = Error (RE + SE)

**RE** = Random Error

**SE** = Systematic Error

## Systematic Error (SE)

SE is caused by factors operating during the test that should (and often can) be controlled. An obvious source of SE is cheating, which increases the test scores of cheaters, but not the scores of the rest of the test takers. This effect is certainly not part of the TS of test takers. Cheating is not random error, as there is absolutely nothing random about efforts to cheat.

Testwiseness is another—equally damaging—source of examinee-based SE. Some test takers are very good at evaluating testing cues, while other test takers lack that ability. Even if testwiseness and cheating affected all test takers equally—which they do not—they would still be sources of SE. They inflate scores,

making them inaccurate, and they lead to mistakes when decisions are made based on test scores.

Systematic errors are the most damaging when they affect only a portion of the test-taking population. This leads to severe unfairness and goes against the *Standards*, which we agree to uphold. Haladyna and Downing (2004) stated that "If the estimation of systematic error is impossible, then test scores may not be validly interpreted and, thus, be of no use to the test sponsor." (p. 19).

## AUTHOR'S NOTE:
## TERMINOLOGY RELATED TO SYSTEMATIC ERROR

The term "systematic" refers to an effect based on the conditions or systems of testing, not on a random variable. Systematic error affects an observed score due to a known or unknown factor. The factor can systematically raise an observed score or lower it. For example, if a student is ill while taking the test, that factor (illness) will likely lower an observed score. Remember that systematic error is always unrelated to and independent of the skill being measured. Systematic errors are difficult to discover, even if one is looking for them.

Systematic error can be consistent across all test takers, producing a systematic effect that is the same for everyone. For example, a scoring error that adds five points to each test taker's score is a consistent systematic error. If discovered—and that isn't guaranteed—it can be quickly fixed. Other systematic errors are inconsistent, applying differently to test takers. An example would be when some test takers have access to pre-knowledge about the exam content, but others do not. Respectively, some scores are much higher than they should be, and others are unaffected. These inconsistent systematic variables are also difficult to discover, and their effect will likely be more complex, making the task of removing those effects almost impossible. Some, such as cheating, have existed for decades, and their effects are still not well-understood, and remain a thorn in the side of testing programs.

There have been several names used as close-enough synonyms for systematic error, and the reader should be aware of them. Systematic variables that lead to systematic error have been called *extraneous variables*, mainly when referring to scientific experiments. Extraneous means "not relevant to" or "not a part of." In testing, extraneous variables are not part of or not relevant to a person's true score (TS) on the skill being measured.

Systematic variables have also been referred to as *secondary variables*, meaning they are not of primary relevance to the measurement task. This does not mean they are unimportant. They are very important because they introduce systematic error, and they make observed scores uninterpretable or useless. Ideally, all secondary variables, extraneous variables, and systematic variables can be discovered and dealt with.

Another synonymous term, specifically from our field of testing, was introduced by Sam Messick (1989). Based on his preference that psychological "constructs" underlie all psychological or educational measurement, he referred to systematic error as *construct-irrelevant variance* (CIV). In other words, systematic error, either raising a test score or lowering it, is due to construct-irrelevant easiness or construct-irrelevant difficulty, respectively. Messick referred to CIV as a "contaminant" with respect to score interpretation. So, systematic variables can also be called *contaminating variables*, probably a more apt label than the others. Messick used the phenomenon of testwiseness as an example of how a construct-irrelevant variance, irrelevant to the construct being measured, can raise test scores.

One final term for systematic variables and systematic error is *confounding variables*. This is closer to the concept of contaminating variables in that their presence confounds or confuses the interpretation of scores. Is the score due to the individual's true ability, or to some other factor?

Whatever we choose to call them, our responsibility toward such contaminating variables is to (1) accept that they exist, (2) make every effort to discover them, and (3) prevent them from affecting test scores. As testing professionals, it's our responsibility to continue to search for and get better at controlling these unwanted influences on our test scores.

## OTHER TERMS FOR SYSTEMATIC ERROR

Secondary Variables

Contaminating Variables

Confounding Variables

Extraneous Variables

Construct-Irrelevant Variance (CIV)

When giving tests, program administrators try their best to "standardize" the test administration and make sure every examinee has an identical experience. For this reason, whenever possible, test takers are given identical content, tests in the same format, identical instructions, and the same time limit. These examples show how program administrators attempt to reduce SE and promote fairness.

Unfortunately, some SE occurs when attempting to standardize test administration. Standardization is unable to eliminate all sources of SE. Indeed, most programs are not able to control all sources of SE or even most of them. Elimination of SE is a noble goal, but rarely realized. And efforts to track whether standardization successfully eliminated all sources of SE are rare. This inability to effectively eliminate all sources of SE introduces variability among test scores.

For example, consider a set of students taking a test on computers; all of a sudden, the power goes out, and the Internet connection fails. For some of those students, waiting for the test to resume is stressful and increases anxiety, which will negatively affect their test scores. Another example of SE is fatigue; some individuals might be more fatigued by a lengthy exam than others. When the skill set being tested in most exams is not endurance (as is the case with nearly all exams), fatigue becomes a source of SE.

These are just two examples of SE that are routinely present in the design, development, and administration of tests. There are many more. SE introduces unfairness when test takers are differentially affected by them. Even in instances when sources of SE affect all scores identically (a rare occurrence), it still impacts the accuracy of the test scores.

## AUTHOR'S NOTE: **SYSTEMATIC ERROR POSING AS RANDOM ERROR**

In the science of testing, as well in other sciences, we tend to call something we can't control "random" error (RE). Here is an example: During a state assessment at an elementary school, a fire alarm goes off during the exam administration. Students dutifully shuffle out of the classroom and return 15-minutes later. The alarm has certainly disrupted the train of thought of every student involved, and has likely negatively impacted the scores for at least some of the students.

Psychometricians treat this type of anomaly as RE, either because they were not aware of the incident, or because they feel it has a minimum effect on test scores. Since its effect isn't known, we treat it as RE so we can calculate reliability, validity, and comparable test scores. However, this error isn't really random; it is simply a small instance of uncontrolled or uncontrollable SE. To the students and teachers affected, however, it is undeniably unfair. Why should they be penalized for such an event when other students at other schools were not? Why can't the students re-take the test under more normal conditions?

## AUTHOR'S NOTE CONTINUED...

Please remember the incidents we usually call out as RE during the course of test development or test administration are almost always actually SE that we are helpless or unwilling to control. It is the responsibility of all testing professionals to recognize the influence of sources of SE and to use whatever means available to eliminate their effects. The introduction of innovative technology into the field of testing holds new promise for controlling this type of error.

SmartItems are not a solution to such errors. Tests made up of SmartItems will be impacted similarly by instances comparable to those described above. SmartItems only nullify the effects of SE impacted by item content or format—but I use the term "only" carefully, as those sources of SE can be very large and very harmful.

### Random Error (RE)

Random error effects are often routinely and knowingly built into test and item designs. One example is the random selection of items included in CAT algorithms to control item exposure levels. Another example is the random assignment of items within content strata to test forms that help create more closely equivalent forms. Other examples include the randomization of item presentation order when items are presented on computerized test forms. The primary goal of each of these effects is increased security.

These randomized design elements are often introduced to make a test form and its items less predictable, or to balance the exposure rate of items. In more general terms, these contribute to removing the effects of the systematic errors, specifically systematic bias from cheating.

These efforts to increase security through certain types of randomization have the direct result of increasing random error during testing. I refer to this as "true random error," as it is generated from an actual random process. This careful use of random error has a wonderful outcome: it removes the damaging effects of cheating. It is a very reasonable trade-off; though randomization adds a small amount of actual RE, it removes large sources of damaging SE.

Fortunately, in addition to all the benefits of removing bias from SE, the added RE is not difficult to deal with. RE has a unique property: with enough observations, it has an expected value of zero. Stated another way, random errors have both positive and negative effects, which balance out with enough observations. Eventually, the amount of error gets smaller and smaller, approaching a value of zero. In the context of this book, the more SmartItems that are given on a test, the closer the RE for a test approaches an expected value of zero.

With enough observations, true random error balances out and the formula can be re-written as:

**Formula 3:**
Observed Test Score = TS + SE + 0

Formula 3 simply states that with enough "observations," RE ceases to have any appreciable effect on an OS. To be clear, all of the values in the formula are theoretical. In actual practice, the values are subject to the variability of estimates.

The questions then become:

**How can we utilize the formula above to best help our true scores and observed scores align?**

**Is it possible to decrease or eliminate some sources of systematic error by purposely increasing random error in tests?**

**Can the randomization of test content and the creation of unpredictable test forms render cheating unlikely or even impossible?**

It seems radical at first, yet this strategy is not unheard of in our industry. The examples given earlier in this chapter illustrate that we have been using randomization successfully for decades in order to make it more difficult to cheat. SmartItems simply take the use of randomization to a more impactful level.

By design, SmartItems introduce random error into a testing experience. This error is not just because of differences in difficulty, but because of other features of item performance, including discrimination, format, length, complexity, etc. It is this characteristic of SmartItems that raises initial concerns from testing professionals, who have a well-deserved distrust of the word "error." Deliberately introducing random error seems to be the opposite of standardization, and entirely against this tenet of testing practices that has been at the foundation of measurement standards for over a century.

However, there is an argument for the increased use of randomization, even though it produces RE. Simply stated, the production of RE is eventually harmless as long as enough SmartItems are administered as part of the test. A single rendering from a single SmartItem is not enough to minimize or eliminate the RE, but with enough renderings, the RE for each test taker will be close to zero.[6] This makes the RE sufficiently small so individual test scores are comparable and can be used for high-stakes decisions.

By adding RE to the measurement process, we can eliminate the effects from some sources of SE. In doing so, our test scores will be more accurate and more trustworthy. With SmartItems, the nefarious impacts of cheating and testwiseness will finally be removed from the testing ecosystem, and a person's test score will more closely represent their ability.

Even when presented with these logical and theoretical rationales for using SmartItems, many remain skeptical of such blatant use of randomization. It is difficult to accept that randomization can have positive and helpful effects when, for decades, we have followed the principle that the standardization of items and tests is inviolate. Even though randomization is increasingly used in most test designs, its use for test content seems to be distrusted, questioned, and discouraged.

However, this doesn't mean our industry isn't willing to embrace small random error when the benefits are clear. As an example, research has shown there are small but noticeable "order effects" from randomly presenting items and options on tests. However, we have continued with this practice because:

1. This random error reaches zero with enough observations, and
2. The randomization of options and items are used to counter serious problems from cheating.

**The trade-off is worth it.** As with randomized item order on tests, the random error introduced by SmartItems can be addressed and controlled, and the benefits are undeniable.

## SmartItems | Practical Arguments

The practical argument for SmartItems is this: Despite its newness, oddness, and requirement that we do things we have never done before, **SmartItems are worth it because they decimate critical sources of systematic error including test fraud and testwiseness**. In essence, the benefits of SmartItems outweigh the costs.

### Testwiseness

Testwiseness is the ability to use cues inherent in item formats, rather than knowledge of the skill being measured to answer questions correctly. Often referred to as "test-taking skills", I would guess most people reading this understand the unfair power of testwiseness. This is either because you possess test-taking skills which have helped you on exams throughout your life, or because you have always struggled with test-taking skills, and this has hurt you on tests throughout your life.

Regardless, test-taking skills are **an accepted, unfair aspect** of testing. Researchers in the 1950s established the phenomenon, and ever since, we have been looking for ways to remove these effects from exams.

### Test Fraud

Effective cheating is probably the most significant threat to the valid interpretation of test scores today. As a founder of a test security company, I am acutely aware of the unfair impact of cheating and theft. My colleagues and I have witnessed the devastating effects of cheating across all areas of testing, including K-12 education, educational admissions, higher education, certification, licensure testing, and employment-based testing. Unfortunately, after 15 years, we have not seen an overall decrease in the prevalence of these two types of fraud. Oh, I believe as an industry we have gotten better at managing test fraud, but **the cheaters continue to get better as well**!

The simple truth is that while the amount of fraud a testing program faces varies, it is always present. More importantly, it's almost impossible for programs to tell the difference between a test score that has been enhanced

by cheating and one that hasn't. As a result, we open ourselves up to the valid criticisms of reporters and other societal watchdogs who cast justified doubts on the decisions being made based on our test scores.

Our traditional approach to test security—which occurs mainly during test administration—is no longer an adequate solution. Something else must be done.

Measurement professionals seem to tolerate and accept both testwiseness and test fraud as necessary evils.[7] However, this does not need to be the case—SmartItems have the potential to eliminate them both. The important question then becomes: "Are SmartItem worth it, even though they introduce some random error?" Maybe it is better to continue to accept the status quo than to disrupt our time-honored processes and efforts at standardization. Perhaps SmartItems are just too difficult to implement in testing programs. Is it better to change current processes, policies, and procedures and explain SmartItems to stakeholders and examinees? Or should we continue to accept the tolerated pain and uncertainty of fraud and testwiseness?

I had a minor stroke recently, which, as the doctors all told me, makes it more likely I'll have a stroke in the future. Without making significant lifestyle changes (e.g., avoiding certain foods, exercising daily, taking medicine as prescribed, etc.), I would live under the constant threat of another stroke. For me, the choice was clear. I did not want to suffer another stroke. The lifestyle changes seemed small by comparison.

The choice to use SmartItems is similar—use SmartItems and accommodate the relatively minor changes they would require, or continue to accept the large systematic errors caused by cheating and testwiseness. For testing professionals sworn to remove sources of systematic error, the choice should be clear.

### AUTHOR'S NOTE: **ERROR'S "LARGE" IMPACT**

What do I mean by "large"? For example, from our Caveon Data Forensics work, we have seen exams where *most* of the scores have been increased by cheating. We have never—not in 15 years—seen a set of clean test results that are completely unaffected by cheating.

The best-case scenario today is that "only" 5% to 25% of test scores have been increased by some form of cheating. Regarding testwiseness, scientific research using DOMC indicates that when traditional multiple-choice questions are used, test scores are raised 5% to 10% by using testing cues totally unrelated to the skills being measured. These amounts can be considered very large, and should be unacceptable to testing professionals.

## SmartItems | Empirical Arguments

The question about SmartItems I most appreciate hearing is, "Do they work?" The question, while only three words, is multi-faceted. Here are some ways of re-stating it:

- Do SmartItems contribute to reliable exams?
- Do SmartItems contribute to making valid interpretations from exam scores?
- Are SmartItems fair?
- Do SmartItems actually prevent the theft of test content?
- Do SmartItems actually prevent cheating?
- How easy are SmartItems to use?
- Do you have any proof of these claims about SmartItems?

These questions are about the practicalities of SmartItems. With SmartItems, **will we be able to achieve safer, more actionable test scores**? What evidence exists to answer that question?

The empirical argument for the use of SmartItems is this: we have designed, run, and collected data from research projects that "prove"[8] SmartItems work as promised. The results are very encouraging. **While more research is always welcomed, there is enough empirical evidence today to encourage the serious consideration of SmartItems.**

A detailed description of the several research projects and simulations that have directly evaluated SmartItems is given in Chapters Seven through Ten.

1. SmartItems perform well psychometrically. Statistical analyses indicate SmartItems perform in the range where psychometricians would support their use.
2. Tests made up of SmartItems demonstrate high levels of reliability.
3. Tests made up of SmartItems demonstrate high levels of validity.
4. SMEs report being capable of (and even enjoying) the task of designing and building effective SmartItems.

These results are surprising and almost unbelievable to most testing professionals. It is hard to believe that **SmartItems, varying as they do, can be described using common item analysis statistics**. It is even harder to believe they can contribute to test reliability and validity.

## AUTHOR'S NOTE: RESEARCH ON CURRENT TESTING METHODS

It is probably fair to say that more research has been done on SmartItems than on many other technology enhancements currently used for testing, including:

- The use of the drag-and-drop functionality in item design
- The comparison of the select-all-that-apply multiple-choice item type with other item types
- The utilization of online proctoring for exams
- The effectiveness of authentication and identification methods
- The comparison of devices used for testing, such as smartphones and tablets

These technology enhancements are widely used in our industry, yet rarely studied.

# 03 CONCLUSION

SmartItems achieve their promise. Not only do they bring tremendous advantages; they actually work! SmartItems are not something to be feared. There is enough empirical evidence—along with logical, theoretical and practical arguments—to encourage testing programs to move forward with preliminary discussions, with trying out the technology, with conducting pilots and field tests, with performing actual scientific research, or with implementing SmartItems as a solution.

I hope these arguments answer your questions about how SmartItems can be fair when two test takers might randomly see items of varying difficulty. Rather than threaten fairness, SmartItems actually enhance it by eliminating the two largest sources of unfairness: testwiseness and cheating.

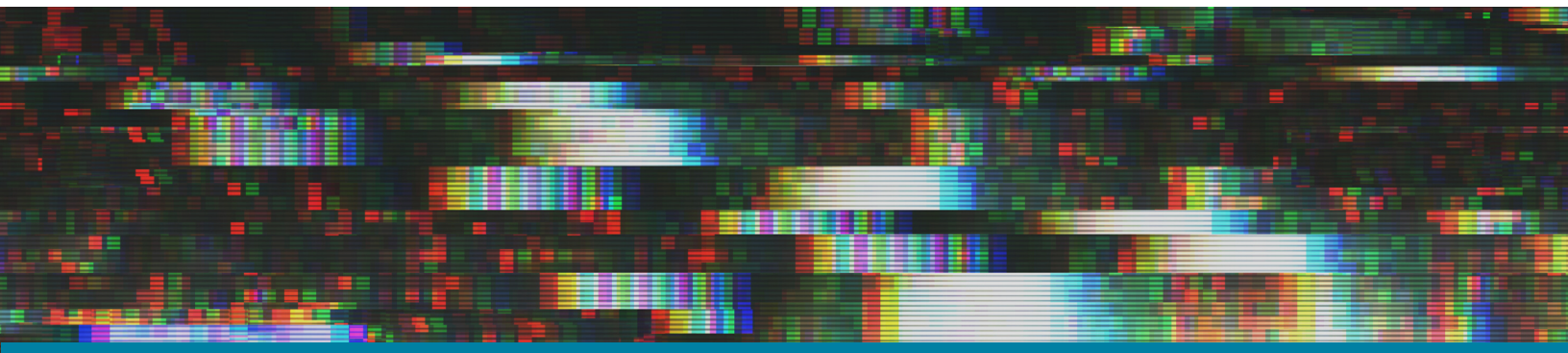
So, rather than asking whether SmartItems are fair, a more important and optimistic question might be: "How do tests made up of SmartItems produce comparable and fair test scores?"

## FOOTNOTES:

1. I have always found it interesting that the questioner always refers to "difficulty" as the way SmartItems differ, when there are many other ways items can differ, whether or not randomization is involved. And all of these affect how "fair" the item is. These include discrimination, reading level, length of time to process the item, susceptibility to guessing, susceptibility to testwiseness, and others. The randomization of SmartItem renderings will have a similar impact on these less considered differences as it does on difficulty.
2. This refers to the *Standards for Psychological and Educational Testing* published by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The latest version was released in 2014 and contains an entire chapter devoted to fairness, an important component of validity.
3. This argument assumes items are differentially difficult for individuals with higher or lower ability on the skill being measured.

## CONTINUED:

4. Testing terminology: A "form" is a version of a test. Often, the term used is "test form." For example, a test, for security purposes, may consist of one or more forms. If there are two forms, Form A and Form B, it doesn't matter if a test taker sees one or the other. Scores from both test forms are comparable or can be made comparable by statistical adjustment.
5. Most tests will not consist entirely of the same SmartItem duplicated multiple times. The typical high- or low-stakes tests will have many different SmartItems, each one covering a different specific skill as part of a larger domain. For example, a test of sixth grade math will have a SmartItem for each standard the test is designed to cover. It is conceivable that for those skills that are more highly weighted, the SmartItems can be duplicated to reflect the weighting.
6. Simulations of the effects of random presentations of items are presented in Chapter Seven. These simulations indicate that a reasonable number of SmartItem renderings will produce comparable test scores.
7. The circumstance with both of these reminds me of the phenomenon "learned helplessness," which I learned about as a graduate student. Learned helplessness is an effect caused by repeated exposure to adverse conditions while at the same time not having any ability to escape or avoid them. The result of this repeated exposure is a personal acceptance of the conditions along with the pain that comes with them. After a period of time, no more attempts are made to fix the causes.
8. The word "prove" is used casually here to mean several actual scientific studies and case studies have been conducted which support the claims made about SmartItems. In reality, scientific and quasi-scientific research is never able to "prove" hypotheses, but simply able to reject competing hypotheses with a particular level of surety.



Chapter

# 04 ERRORS, ERRORS, EVERYWHERE

## DEALING WITH COMMON, SERIOUS, SYSTEMATIC ERRORS IN MEASUREMENT

This chapter acknowledges two intractable sources of systematic error, and suggests ways of addressing them. While SmartItems are an effective solution for eliminating these sources of error, I will also discuss other alternatives available to the testing community.

The chapter takes its direction from the methods scientists use when they run experiments. Scientific experiments are as much about controlling extraneous variables as they are about discovering cause and effect. As such, they represent an effective model for our own efforts to control error.

Throughout the chapter, it is important to keep in mind the two sources of systematic error introduced in Chapter Three: test fraud and testwiseness. Doing so will help us realize that even though SmartItems introduce random error, the trade-off is worth it. Random error is relatively easy to address and deal with, and the benefits of eliminating fraud and testwiseness cannot be overstated.



In Chapter Three, we discussed some fundamentals of test theory. Whether "classical" or "modern," these theories can be summarized in eight points:

## TEST THEORY FUNDAMENTALS

1. Observed test scores are inaccurate and affected by two sources of error; random error (RE) and systematic error (SE).

2. RE is unpredictable and unbiased. It can have a positive or negative effect in equal likelihood. With enough observations, RE has an expected value of zero.

3. Systematic error, also called construct-irrelevant variance and other names, is biased, difficult to spot, difficult to deal with, and sometimes unknown. For these reasons, it is the more dangerous type of error. Cheating and testwiseness are the common and nefarious examples of systematic error that are the focus of this book.

4. Test theorists acknowledge systematic errors and encourage efforts to discover the source of such errors and control for them. Unfortunately, we're offered little in the way of practical advice on how to actually do this.

5. Test theory, whether classical or modern, is heavily dependent on random error being the only acceptable source of error affecting test scores. As such, these theories only support our efforts as testing professionals if we control systematic error.

6. Systematic error is a greater threat to the validity of our exams than random error. If these sources of error are not found and managed, it is impossible to use test scores validly or dependably in high-stakes decisions. Today, many of these high-stakes decisions are made in ignorance, since the decision-makers involved are often unaware of compromised tests, or either ignore the evidence from frequent security breaches or disregard the implications and common criticisms of multiple-choice tests.

### TWO IMPLICATIONS CAN BE DISTILLED FOR SMARTITEMS

7. Increasing random error (as SmartItems do) is not harmful if enough items are included on the test. With enough items, the effect on test scores from random error is virtually nil.

8. If we can control and eliminate systematic errors by increasing random error, then the overall impact on the validity of test scores will be positive.



Discovering sources of systematic error is not particularly difficult. Cheating is both common and easy to detect with the right security processes in place, and the sources and consequences of testwiseness have been documented for 70 years.

Indeed, if you trust research and media reports, you can't swing a stick without accidentally hitting someone who has cheated or who regularly uses finely-honed testwiseness skills. These two sources of systematic error are already well known and accepted, and they continue to confound every test score on every test—particularly those with multiple-choice questions.

It is true testwiseness and test fraud are not the only sources of systematic error. Other sources can develop from the characteristics of a person, a group, a test, a testing location, a scoring procedure, and others. However, most sources of systematic error are more subtle than cheating or testwiseness, and others still may be unknown. Systematic error can arise because of the adoption of technology, or because we continue to use traditional methods and tools. Unknown and little-known sources of systematic error will take more effort to discover, but it is a trainable and worthwhile skill for all testing professionals. It should be standard instruction in every measurement course. Some sources of systematic error identified by other testing professionals are listed in Figure 4.0.

**Figure 4.0.** Sources of Systematic Error.

### Cheating
Researchers: Haladyna and Downing (2004); Downing and Haladyna (2004)

### Insecure Content, Item Theft
Researchers: Downing and Haladyna (2004)

### Testwiseness
Researchers: Haladyna and Downing (2004); Messick (1989); Thorndike (1951)

### Lack of Reading Ability
Researchers: Messick (1989); The *Standards*; Downing and Haladyna (2004)

### Motivation
Feldt and Brennan (1989)

### Test Anxiety
Researchers: Haladyna and Downing (2004); The *Standards*

### Fatigue
Researchers: Haladyna and Downing (2004)

### Testing Conditions
Researchers: The *Standards*; Feldt and Brennan (1989)

### Item Formats
Researchers: The *Standards*; Downing and Haladyna (2004)

**Figure 4.0** Continued.
Sources of Systematic Error.

### Biased/Poorly-Written Items
Researchers: Downing and Haladyna (2004)

### Rater Severity in Scoring
Researchers: Haladyna and Downing (2004): Feldt and Brennan (1989)

### Improper Scoring, Cutscore
Researchers: Downing and Haladyna (2004); Haladyna and Downing (2004)

### Teaching to the Test
Researchers: Downing and Haladyna (2004)

### Test Form Difficulty
Researchers: Haladyna and Downing (2004); Kane (2010)

### Test Design
Researchers: Kane (2010); Kane (2011)

### Equipment Used to Measure
Researchers: Feldt and Brennan (1989)

### Biased Tests
Researchers: Kane (2010)

### Standardization
Researchers: Kane (2010)

One problem with many sources of systematic error is that they can be unknown. Also, even when sources of systematic error are well-known, they can still go unaccounted for because they are difficult or even impossible to measure. Here are two examples:

1. The use of pre-knowledge by a test taker to cheat is virtually undetectable, whether searching during the exam or after.

2. It often isn't clear which students are better at taking tests than others—most can remain calm under time constraints, and many are trained to efficiently compare and eliminate unlikely options.

Many personal characteristics are sources of systematic error. The best method for controlling them is not through the daunting task of trying to evaluate the amount of error contributed by each individual, but through test design. The use of SmartItems (especially SmartItems enhanced by the DOMC item type) will easily eliminate not only testwiseness but many of the other sources of systematic unfairness (see Figure 4.0).



## Science and the Control of Systematic Error

Let's take a moment to revisit how scientists control systematic error in scientific experiments. Scientists, when conducting experiments, specify two types of variables. One is called the Independent Variable (IV); the other is called the Dependent Variable (DV). The IV is the variable the scientist will manipulate in order to see if changes in the IV will affect the DV. The Dependent Variable is not manipulated. Instead, it is observed, and any changes are recorded. In any scientific experiment, every effort is made to control confounding variables that might impact the results.

### AUTHOR'S NOTE:
### UNDERSTANDING THE DIFFERENCE

What's the difference between IV and DV? Imagine a botanist studying the effects of different fertilizers on plant growth. The different fertilizers represent the IV, and the amount of growth is the DV. The more effective fertilizers would result in greater plant growth. Or, imagine a researcher is studying how various artificial sweeteners affect the preference for specific soft drinks. The IV is the sweeteners, and the DV is the stated preference for soft drinks containing those sweeteners.

In all experimental studies, scientists are concerned with making sure no extraneous variables impact the results. The botanist would want to make sure all plants were watered similarly; the food scientist would want to make sure participants were equally thirsty at the time the study was conducted. Members of the testing industry should consider the way scientists control systematic error to help increase the overall confidence in an experiment's results. We should be able to apply similar methods to our approach to testing.







To be clear, this discussion of best practices in scientific experimentation is not an analogy or a metaphor; it is a direct illustration of control techniques used to address unwanted, contaminating, or extraneous variables that might confound the interpretation of an experiment's results. One might think, "This directly translates to our goal of eliminating variables that confound the interpretation of test scores." That observation would be correct. The tools are the same, though the purpose in measurement is slightly different.

I was trained as an Experimental Psychologist and spent many painstaking hours learning the proper methodology for conducting experiments. One of my professors, Dr. Paul W. Robinson, authored a popular book on research methods that was (of course) required reading for the introductory research methods class (Robinson, 1981). I not only learned from that book, but as a graduate student and professor teaching the course in later years, I used it for my students. To this day, I still recall a specific section that stressed the importance of controlling any extraneous variable when conducting an experiment. Without controlling the extraneous variables, results might have a confounding and alternate explanation.

Let me describe a fictional experiment involving human subjects: Imagine an experimenter wishes to know if training on a new problem-solving method will help high school students solve common problems. The experimenter sets up a design with two groups of students. One group is trained in the new method, and the other is trained on a neutral topic unrelated to problem-solving. This "training" is the IV. Before and after the training, each subject will be given ten problems to solve, and will get a score out of ten each time. The difference between the before and after scores is the DV. If the training works as hoped and increases problem-solving abilities, the researcher expects to see the scores increase after training. This would demonstrate that changes in the IV caused the changes in the DV

However, there is one important step that must take place before the experiment begins. The experimenter must make sure all extraneous variables are addressed and managed to more clearly show that the IV caused the changes in the DV.

The students will be divided into two groups, but the experimenter must make sure the groups are equivalent on the most important potentially contaminating variables. For example, age might be important. It is likely older high school students would already have more experience in solving problems than younger students would. Another example might be grade-point average (GPA). Those who have excelled in school courses may solve the problems better than those with lower grades. The experimenter lists a large number of other possible confounding variables, including gender, socioeconomic status, language, cultural differences, and more. The goal is to make sure each group is equally represented with regards to all of these variables.

According to Robinson (1981), there are several methods available for controlling these variables. These are outlined in Figure 4.1, along with a description of the control method, terms, and examples for how this method can be used for the same purpose in testing.

**Figure 4.1**. Scientists' and Test Developers' Control of Systematic Error.

**SCIENTISTS' AND TEST DEVELOPERS' CONTROL OF SYSTEMATIC ERROR**

| CONTROL METHOD | SCIENTIFIC DESCRIPTION | TESTING DESCRIPTION |
|---|---|---|
| CONSTANCY / ELIMINATION | Eliminate the effects of extraneous variables by holding them constant.<br><br>*Example: Our researcher could ensure all subjects in the experiment were equivalent for a given variable (e.g., use only subjects whose primary language is English, or only subjects who are male). With constancy, the influence of those extraneous variables on the dependent variable is completely removed.* | In testing, we call this method standardization.<br><br>*Example: We attempt to ensure the conditions of testing are constant for all test takers (e.g., using the same test form for everyone, making sure everyone has the same time limit, or giving the same instructions to the test under similar distraction-free conditions).* |

**Figure 4.1** Continued. Scientists' and Test Developers' Control of Systematic Error.

| CONTROL METHOD | SCIENTIFIC DESCRIPTION | TESTING DESCRIPTION |
|---|---|---|
| MATCHING | Make sure extraneous variables were represented equally in both groups<br>.<br>*Example: Our researcher could ensure each group of students was made up of the same number or proportion of a given variable (e.g., the same proportion of male and female students or English speakers and non-English speakers). With matching, the influence of the extraneous variables is not eliminated, but spread evenly in the two groups.* | This control method is not commonly used in the field of testing.<br><br>*Perhaps the closest example of this is when test forms are balanced on the basis of content. Test developers wish each test taker to get the exact same cross-section of the content; otherwise the imbalance will affect test scores.* |
| STATISTICAL CONTROL | Modify results statistically by removing the effects of a measured secondary variable.<br><br>*Example: Our researcher could measure and record a relevant but extraneous characteristic of the individuals assigned to each group (e.g., GPA). The researcher would then statistically remove the effects of GPA differences from the DV.* | Statistical control is used often in the field of testing.<br><br>*Example: When equating test forms, we first measure the performance of each form, and if one form is found to be easier than a second form, the scores from the harder form can be adjusted upward by the proper amount. In doing so, any systematic error resulting from differences in form difficulty is countered.* |

**Figure 4.1** Continued. Scientists' and Test Developers' Control of Systematic Error.

| CONTROL METHOD | SCIENTIFIC DESCRIPTION | TESTING DESCRIPTION |
|---|---|---|
| TURNING AN EXTRANEOUS VARIABLE INTO AN INDEPENDENT VARIABLE | Consider focusing research attention on a secondary variable. This requires measuring it by changing the experimental design and reporting its influence.<br><br>*Example: Our experimenter could control an extraneous variable by making it an independent variable and studying it directly (e.g., our experimenter could study whether there are gender differences in the use of her problem solving training by dividing each of the two groups into male and female subjects, making four groups in total).* | There is no clear analogue for this control technique in the regular day-to-day world of test development and test administration.<br><br>*Example: It could exist in the academic research on testing. Using the extraneous variable of cheating, we could divide the test takers into two groups: those who are allowed to cheat on the test and those who are not.* |

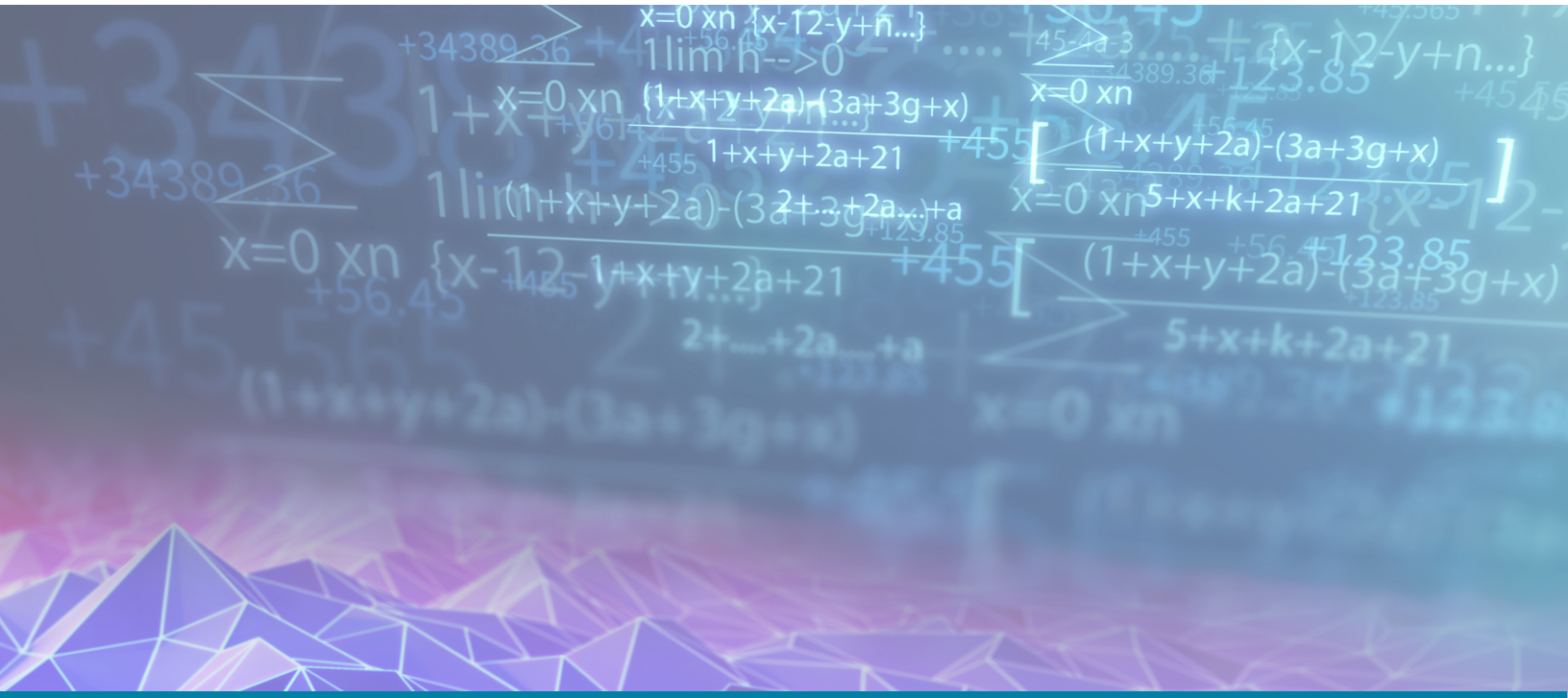


**Figure 4.1** Continued. Scientists' and Test Developers' Control of Systematic Error.

| CONTROL METHOD | SCIENTIFIC DESCRIPTION | TESTING DESCRIPTION |
|---|---|---|
| RANDOMIZATION[1]<br><br>Considered by most scientists to be the ideal solution for dealing with extraneous variables, randomization is viewed as the "gold standard" of control methods for scientific experimentation. | Randomization in experimentation almost always refers to the random assignment of research participants into the groups.<br><br>*Example: Our scientist could take the available sample of high school students and randomly assign them to the two groups (e.g., by flipping a coin or using a random number generator on a smartphone). Randomization will make sure the two groups are practically equivalent on ALL known and—just as important—unknown extraneous variables.*<br><br>There are two limitations of randomization. First, the sample size has to be sufficiently large. Second, randomization isn't always possible in scientific studies, especially those conducted in the field rather than in the lab, as it is not often straightforward to assign research participants to independent groups. | Randomization is often used in the field of testing.<br><br>*Example: Randomization in testing is typically used to counter an extraneous variable associated with test security (e.g., item order is increasingly being randomized in computerized tests. This makes it difficult for test takers to use a pre-published "key," as such documents rely on a fixed order of items).*<br><br>In addition to the current ways of utilizing randomization, it is possible to also randomize the content of each item in order to make it unpredictable, increasing security by making certain types of cheating and theft impossible. |

More than one of these methods can and should be used at the same time. Take stratified randomization, which is a combination of randomization and matching. Imagine the researcher matched the two groups on gender, having an equal number of males and females. With that variable equalized, each male or female is then randomly assigned to one group or the other. SmartItems can be used as a form of stratified randomization of test content because randomization of content occurs within each SmartItem. If there are 20 different SmartItems on the exam, randomization is stratified by the 20 skills the test is measuring.

In scientific experiments, the ideal method for controlling extraneous variables is randomization. This is particularly true for those variables that pose the greatest risk of confounding the results. It is also the case for any unknown variables the research was unable to locate, predict, or account for. As shown in Figure 4.1, randomization is not the only method for controlling known extraneous variables—but the other four methods have some distinct drawbacks. These include:

## Effectiveness

Other control methods are often unnecessary because it is unclear whether the variables they are meant to control will influence the study results or not. With randomization, these control mechanisms are often rendered superfluous.

## Scope

Other control methods can be expensive and difficult to use, and are usually only effective in controlling one or two extraneous variables. If there is more than one extraneous variable that needs to be controlled, then randomization is easier, cheaper, and more effective.

## Influence

Other control methods limit the generalizability of experimental results. For example, if our researcher decided to use only males in the study in order to control the effects of gender differences, the results would apply only to males. Randomization has the advantage of being able to support the study's ability to generalize broadly to the population from which the sample was drawn.

## Oversight

One final problem with control methods compared to randomization is that the sources of systematic error need to be known in order to be controlled. When there are sources that are overlooked, new, unexpected, or simply unknown, their effects will not be removed from the results, and randomization will remain the best choice.



For all the reasons listed above (and more), randomization is science's preferred choice for controlling unwanted, extraneous, and confounding variables. With that said, there are also some cases where it is possible, and often desirable, to use more than one control method at a time in an experiment. Of course, there are also times when randomization is not possible, particularly in field experiments. In those instances, the other control methods should be applied.

Given the powerful ability of randomization to control for the effects of unknown and unwanted variables, it makes sense to consider making greater use of randomization in the field of testing.

## Testing Today: Applying the Methods for the Control of Systematic Error

Extraneous variables are a plague to the high-stakes testing world. Two examples of these variables, cheating and testwiseness, have been used extensively throughout this book. But the testing industry's attempts to control cheating have been largely ineffective, and our world at large has seemingly accepted testwiseness as a problem better ignored than managed.

Unfortunately, whether we ignore them or not, cheating and testwiseness still account for large amounts of systematic error and severely contaminate our test scores and the statistics routinely used to evaluate our items and tests[2]. Rather than effectively addressing these issues, most of the time we proceed as if these



significant sources of error do not influence our tests. This is a dangerous approach. As Kane (2010) stated, "If systematic errors cannot be removed, they tend to be more troublesome than random errors of the same magnitude because they do not cancel out in the long run" (Kane, 2010, p. 13).

Like cheating and testwiseness, some sources of systematic error are part of the many personal and inherent characteristics of test takers. Unfortunately, since we usually cannot measure them well, we cannot directly control them. If we can't control them, what can we do with them? The answer is simple: We can design our tests better to neutralize these errors.

In the same way it benefits scientists conducting experiments, randomization should be our "gold standard" method for controlling extraneous variables—in particular, those that test takers bring with them to the test. While we already use randomization in limited ways, such as randomly organizing item order on a computerized test, its use is sometimes criticized. However, the fact that the order of

items does have a small demonstrated effect does not mean the practice should be eliminated. While it introduces a bit of error as a random order effect, it more than makes up for it in removing the consequences of errors from cheating. Here are a few more examples of how randomization is used to control extraneous variables in testing today:

**1** Test forms are often randomly assigned to test takers. The security benefits of randomly assigning test forms are:

- Reduced exposure rate of items, and
- Reduced effects of pre-knowledge from stolen forms.

**2** Elements of randomization are routinely applied as part of our CAT item selection algorithms. For example, for the algorithm of an adaptive test I designed in the 1990s, items of a particular difficulty level would be selected from all those that were available in the pool. Once a test taker's ability was re-estimated, and another item needed, an item was then selected randomly from the five most informative items in the pool. One reason for this randomization was to reduce the negative effects from the overexposure of items.

**3** The order in which items on computerized tests are administered is often randomized. This eliminates the ability to create a stolen key of test questions and their answers (e.g., The answer for Question 1 is B). One other reason for randomizing question order is to prevent an examinee from actively copying answers from a neighbor's test.

**4** Randomization can be applied to item content as well. Psychometricians working on AIG projects are agreeable to the use of randomization of surface or incidental information. Although the items differ in exact content, it is assumed these types of random, surface-level changes keep the psychometric properties (i.e., difficulty, discrimination, etc.) equivalent and predictable. These items can be used interchangeably on test forms, without pre-testing and reviews from SME's. Randomization is a critical component of the AIG process.

Examples (Links)

*Linear Syllogism Generator*

*Package 'AIG'*

**5** The option order for computerized multiple-choice questions is often randomized. One reason for this randomization is to make it difficult for item harvesters to create a reliable key. Other non-security reasons include reducing the undesirable effect that happens if the correct answer is always A and eliminating an item writer's bias to place the correct answer in specific positions.



A review of these other uses of randomization in current test and item design reveals that improving test security seems to be the dominant driver of these actions. Efforts for reducing exposure, expanding item pools, and re-ordering items and their options all show that the testing industry is already willing to utilize randomization to foil test fraud. Not at all coincidentally, that is the primary motivation for using randomization in SmartItems.

## Randomization and SmartItems

When is it okay (and even desirable or imperative) to present items with randomized content to test takers?

You have learned about some of the systematic errors that interfere with the interpretation of test scores. As written in the *Standards*, it should be our primary goal as an industry to eliminate the errors from cheating, testwiseness, and other sources of systematic error, and to give our test takers the best possible opportunity to show off their skills and knowledge. As stated in the *Standards*, "A test that is fair...reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the same population" (p. 50). Logically, a score from someone with pre-knowledge of all or some of the test content is not measuring the same construct as a score from a test taker who has no prior knowledge of the content.

According to our industry standards, it is okay to use randomization on a test (including random content) if doing so will remove the systematic error effects of unfair extraneous factors, such as cheating and testwiseness. Removing these effects allows all test takers to better demonstrate their skills and knowledge.

SmartItems randomize content during the presentation of an item. While randomization has been used routinely in other ways during testing, this is the first time it has been proposed to randomize item content every time a test taker sees an item. This is a defining feature of a SmartItem (see Chapter Two), and such randomization certainly introduces some random error.

Keep in mind that it's not just the information (text, images, etc.) presented in the question that can be randomized, but also the way the test taker should respond. For example, a SmartItem could render by changing whether the test taker is given a short answer or a matching item. The design of the random nature of an individual SmartItem depends wholly on the skill description from which it was designed.

There is no doubt that the various renderings of a SmartItem will vary in difficulty. We should not focus solely on the "difficulty" of each rendering, but also remember that

randomization affects other characteristics as well. The renderings will randomly vary in other ways, depending on the skill being measured, including:

- the time to read and complete each rendering,
- the type of response required for each rendering,
- each rendering's ability to discriminate between individuals of different abilities,
- the reading level of each rendering,
- the test taker's ability to provide the answer, etc.,

It is also likely that randomization will introduce unknown differences among SmartItem renderings. For example, differences will result from how specific individuals interact with item content.

Imagine a SmartItem is designed to measure a person's knowledge of human anatomy. One test taker sees a rendering about the ventricles of the heart. This test taker happens to be the daughter of a heart surgeon and has been taught about the heart since she was young. This rendering will likely be easier for this test taker than for others, even if they all prepare equally for the exam.

The randomization of both known and unknown factors gives amazing internal and external validity to an exam. It's a surprisingly simple control technique over extraneous and contaminating factors, and it's a great alternative to our current attempts to standardize every possible testing condition.

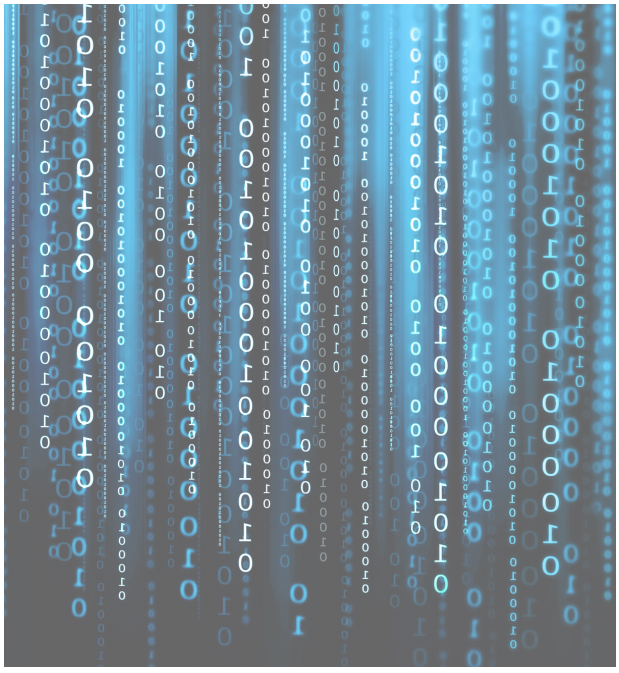
Randomization through SmartItems increases random error, but the good news is this random error will eventually balance out over the course of an exam.[3] The random error effects of randomization end up being small, unbiased, and fair.



## Randomization or Standardization? Is It a Choice?

The randomization inherent in SmartItems does not replace the need for other forms of control over systematic errors in testing. This includes standardization (or constancy). SmartItems mainly affect the presentation of item content. Testing instructions, conditions, locations, devices, security identification procedures, proctoring, environment, etc. should still be controlled. This standardization of much of the test administration process remains important.

Kane (2010) states, "Standardization does have some limitations; in particular, it can introduce systematic errors (by fixing conditions of observations) as it reduces the random error. For example, if some students tend to do better on some kinds of tasks than on other kinds of tasks in the same content area (e.g., objective items vs. essays) may give some students an advantage (positive systematic error) or a disadvantage (negative systematic error)" (p. 13).



## Randomization's Effect on Test Scores, Theoretically

With SmartItems, each test taker will see different items, in a different order, with different content. While many psychometricians object to this use of randomization, there is a general consensus among theorists that with enough items, the random error eventually balances out and has a negligible effect on test scores. In the language of statistics, this negligible effect is sometimes referred to as "statistical error" or "sampling error." Overall, it is considered relatively harmless, especially when compared to the damaging effects of other systematic errors testing programs deal with every day.

On the point of describing random error, Kane (2011) writes:

"Errors that have an expected value of 0 and are uncorrelated with each other and all other variables are called random errors… In some ways, random errors are easier to deal with than systematic errors; since random errors tend to cancel out, we can control them by averaging over more observations… If systematic errors cannot be removed, they tend to be more troublesome than random errors of the same magnitude, because they do not cancel out in the long run." (p. 14-16)

Stanley (1971) states it similarly:

"The errors are random from form to form. They have a population mean of zero for each person. They covary across forms zero with a person's true score and also zero with each other from form to form for a given person." (p. 373)

Likewise, from Embretson (1999):

"Applying the classical test theory model requires several assumptions about error distributions; namely, errors are normally and uniformly distributed in persons, have an expected value of zero, and are uncorrelated with all other variables." (p. 5)

Other sciences have similar theoretical assumptions about random (and systematic) error. These quotes are taken from the Matrix Education (2019) website explaining systematic and random error in taking physical measurements:

"Random errors will shift each measurement from its true value by a random amount and in a random direction… A true random error will average out to zero if enough measurements are taken and averaged (through a line of best fit)."

From Wikipedia (2019) referring to scientific measurements in general:

"Random error is always present in a measurement. It is caused by inherently unpredictable fluctuation in the readings of a measurement apparatus or in the experimenter's interpretation of the instrumental reading. Random errors show up as different results for ostensibly the same repeated measurement. They can be estimated by comparing multiple measurements, and reduced by averaging over multiple measurements."

From a website on chemistry measurement explaining errors of measurement:

"Random (or indeterminate) errors are caused by uncontrollable fluctuations in variables that affect experimental results… A sufficient number of measurements result in evenly distributed data scattered around an average value or mean."

True random error, as opposed to systematic error, is introduced in tests in two ways:

1. Purposefully: It is deliberately introduced by varying testing conditions using randomization.
2. Inadvertently: It is unintentionally introduced from unknown and unpredictable sources.

Either way, random errors are unbiased by definition, helping or harming randomly. With "enough observations," they have no significant effect on scores. While it still needs to be determined what is meant by "enough observations," Toton, Williams, Foster, Foster & Foster Green (2018) reported the results of a simulation that indicated SmartItem tests of fairly short length (~40) might be sufficient. (See more detail about this simulation in Chapter Nine.)

The discussion of randomization makes it clear it is acceptable, desirable, and sometimes imperative to introduce random error into tests, so long as you plan for "enough observations." In fact, using randomization to reduce or remove systematic error should be the "gold standard" for error control mechanisms in all areas of assessment. Test theory supports this tactic, as do best practices of scientists conducting experiments in all other fields.

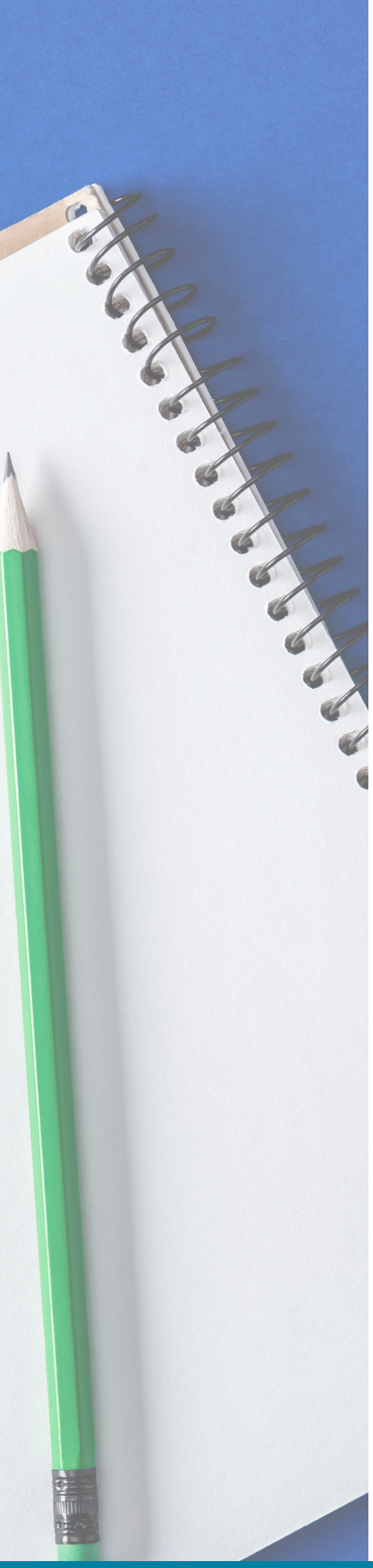
## AUTHOR'S NOTE: **TESTWISENESS IN ASSESSMENTS**

Psychometricians who already use randomization in the limited ways described above should seriously consider expanding its role in tests. There are sources of bias that are not controlled today by standardization and equating efforts. In fact, it can be argued that standardization creates the conditions in which these sources of bias thrive. Testwiseness, an inherent form of score bias supported by the use of multiple-choice questions, is a longstanding, not-yet-dealt-with problem accounting for a 5% to 10% increase in privileged examinee's scores. As long as we continue to use multiple-choice questions, the testwiseness bias will continue to lead us into making poor decisions based on test scores. The random and sequential presentation of options by using the DOMC format eliminates most testing cues associated with the simultaneous presentation of options inherent with the multiple-choice format.

Randomization is a familiar tool for psychometricians, and it can solve these problems in a simple and effective way that standardization and equating cannot. By randomizing item content in addition to our current standardization methods, test developers can prevent the damaging effects of cheating and testwiseness.

## Randomization and Generalization

I'll conclude this chapter by describing one additional benefit of SmartItems due to the randomization of content. In scientific studies, the use of randomization makes it easier to generalize the results of the study to other populations, settings, etc. In areas of scientific experimentation, this is referred to as external validity, or "how valid the results are when placed outside the experimental setting" (Robinson, 1981, p. 113). When extraneous variables are controlled using constancy or matching in scientific studies, these control mechanisms restrict the participant sample, limiting the ability to generalize the results to only those populations from which samples were studied. For example, if a researcher holds gender constant in the study, then any finding from the study will generalize only to that specific gender.





Similarly, it would be valuable if a test score could generalize beyond the specific fixed items that were authored, viewed by test takers, and answered. This is not possible using our current methodologies, where test takers only answer static questions, and therefore the test score only applies (or is externally valid) to those very specific bits of knowledge and behavior. With SmartItems, the game changes significantly.

Because SmartItem renderings are randomly drawn from each SmartItem, and each SmartItem covers the entire breadth and depth of the skill being measured, the score on a single rendering can be generalized to cover the full extent the skill. If a test covers all of the skills for a defined domain, then the test result—the test score—reflects capability across the domain. Tests based on SmartItems indicate a student has learned all of the material assigned, or a worker can complete all of the steps of a complex task. Confidence in a person's ability is enhanced because the test score was produced by an exam free from the common, irrelevant, and confounding variables present in our tests today.

Test theorists, both classical and modern, assume that tests be random samples of and representative of a "universe of items."

Stanley (1971) describes the desirability that test forms be built randomly and independently. He goes on to state: "In practice, no infinite population of forms from which to choose exists, so that independence and random selection are assumed rather than operationally assured." (p. 373)

In 1971, Stanley was not aware computer technology could be capable of creating items that covered entire domains—items that could randomly generate renderings of themselves. With many of these on the same form, his concept of an "infinite population of forms" could be "operationally assured."

Nunnally and Bernstein (1994) prefer to use the term "domain of items" rather than "universe or population of items." In describing the issue of sampling from a domain of items, they state:

"One practical problem is that test items are usually *composed* rather than *sampled* [emphasis added] from a well-defined domain... The purpose of any particular measurement is to estimate the actual random sampling that would be obtained if one could employ all of the items in the domain (e.g., all possible four-digit addition problems)" (p. 217).

Nunnally and Bernstein argue that an estimate of a domain score is obtained from tests consisting of random samples of items sourced from a very large (finite or infinite) pool of items covering the domain. They give an example of a domain as the entire set of four-digit addition items where there are $(10^4)(10^4)$ possibilities, which calculates to 100 million possibilities.

Commenting on the same issue, Crocker and Algina (2008) state, "In the example, it would be impossible to confront the students with all possible problems in long division that they might be expected to solve... Determining the number of items and the variety of content necessary to provide an adequate sample of the behavioral domain is a major problem in developing a sound measurement procedure" (p. 6).

These authors and others recognize the important role of a larger domain of items that, together, represent a single skill or set of related skills. The theoretical concept of a person's true score is based more on the comprehensive population of items than on the set of items ultimately chosen (usually not in a random way) to exist on a particular test. Test scores for individual test takers who took the exam on randomly constructed (whether constructed in advance or on-the-fly from that domain) would better serve as an estimate of their true scores.

Stated another way, scores from tests consisting of SmartItems would have greater external validity than tests constructed using traditional methods. Naturally, those test scores being estimates of true scores, require evaluation by the usual psychometric tools of reliability and validity, as well as evaluating the items with standard statistical analyses.



# 04 CONCLUSION

Randomization is not something to avoid or be afraid of. Rather, its use needs to be expanded and encouraged. Randomization through SmartItems is a method of control that temporarily introduces random error, but has the long-term benefits of preventing the effects of most forms of cheating. In addition, if randomization is used with a DOMC-based SmartItem, testwiseness is also diminished.

The use of recent technology in test development opens up possibilities for better measurement. By utilizing these new innovations and embracing randomization, we can not only remove sources of confounding variables, but make it possible to measure with more relevance, bridging theory and practice more effectively than has previously been thought possible.

## FOOTNOTES:

1. Randomization is often referred to as "random assignment," not to be confused with "random selection." The sample of available individuals would be a better sample if it were randomly selected from the population of individuals. Random selection has little relevance for the discussion in this paper—that of randomization of subjects to experimental groups. But it is very important, as a goal of most (if not all) experiments is to generalize its results to the population from which the sample came.
2. The reader may not appreciate the size of a common testwiseness effect: how much a test score can be affected by some individuals being better at answering multiple-choice questions than others. Studies cited elsewhere in this book demonstrate effects as large as a 10% increase in test scores are not uncommon.
3. If the test has enough "observations" or items. How many SmartItems should be on a test remains a practical question, and one best answered by empirical research.

Chapter

# 05 IMPLEMENTING SMARTITEMS

## THE TECHNOLOGY TOOLS YOU NEED TO GET STARTED

Think of this chapter as a practical guide for using SmartItems on your own exams. If you are inclined to give SmartItems a try and see if the promises are real, congratulations! You have already overcome the biggest barrier to using SmartItems: the unfounded belief that they cannot possibly work. Getting started is a simple process. You just create a plan, integrate some new technology, and explain what you are trying to do for your stakeholders and test takers. This chapter focuses on the second step, integrating SmartItem-enabling technology.

Please be aware that SmartItems are a patent-pending technology. Caveon scientists developed SmartItems according to strict specifications. To that end, a patent on the technology lets Caveon (the only industry organization dedicated to end-to-end test security) make sure SmartItems are implemented properly. To use SmartItems, a testing program must license Caveon's technology. This is a simple and painless process, and one our team is always ready to discuss. Click here to reach out, or go to http://bit.ly/ContactCaveon.



## Software for SmartItems

SmartItems are dependent on technology. They cannot be implemented on paper-and-pencil tests or on computerized tests that are limited to traditional and simplistic designs. They also cannot be used with test development and delivery tools that have not been designed and built to support them. That said, accessing the appropriate technology is relatively simple and inexpensive; programs interested in SmartItems can either switch to a robust software that supports SmartItems, or enhance their existing development and delivery systems by integrating with a SmartItem API. These two processes are outlined in Figure 5.0.

### Option 1:

#### Caveon's SmartItem Software

Utilizing SmartItems is surprisingly easy with Caveon's item development and delivery platform, Scorpion™. This tool natively accommodates the use of SmartItems (including SmartItems that use DOMC) and builds them using three methods:

1. Using a code editor
2. Using an easy drag-and-drop user-interface, and/or
3. Creating an unlimited number of options for selected-response questions.

After SmartItems are built, they are then rendered during test administration. Scorpion collects the necessary data for evaluating SmartItems, allows data export options, and has built-in analyses for quick, convenient feedback.

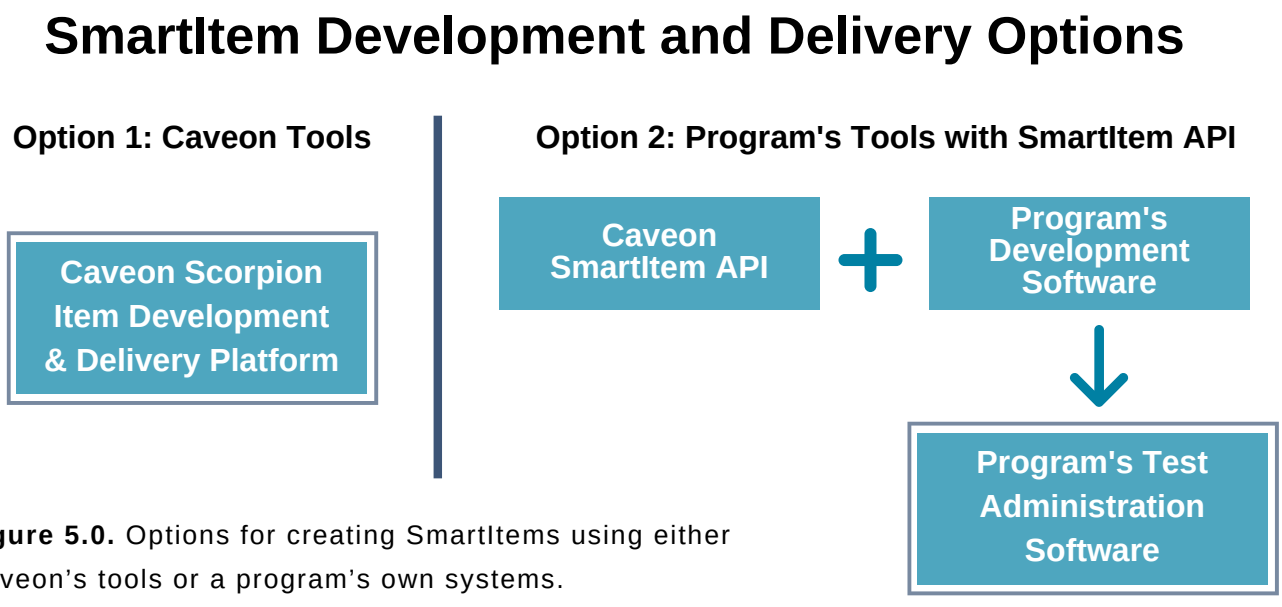



**Figure 5.0.** Options for creating SmartItems using either Caveon's tools or a program's own systems.



## Option 2:

### SmartItem API

For programs fully committed to current test development and test administration software, SmartItems are still accessible. Caveon's SmartItem API can quickly and easily integrate into your existing systems, providing the capability and benefits of SmartItems without having to change testing infrastructure. No rip-and-replace needed. The SmartItem API, when implemented, can be used by your own item writers to easily create SmartItems for any of the item types currently supported by your systems. To further your program's security initiatives, you can also license the DOMC technology, which will eliminate the testwiseness effects inherent with the multiple-choice item format.

To further simplify the process, the SmartItem API is a GUI interface that makes it easy for a SME/item writer to create a SmartItem without actually writing code. Figure 5.1 shows a screen from the SmartItem API, illustrating how variables are created and used in an item. By designing the variables needed, and creating lists of item components, an SME can easily develop a fantastic SmartItem.These two technology options make it possible for any assessment program to unlock the benefits of SmartItems. Moving forward, let's look more closely at the experience of creating these innovative items.

## Choosing your SmartItem Method

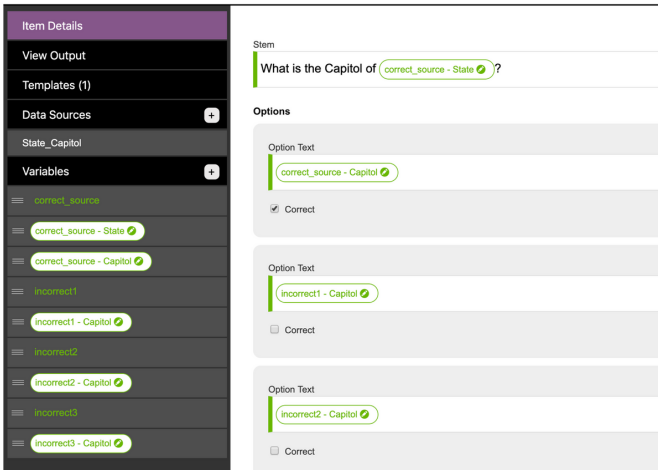
Many programs are concerned that their SMEs will need to know how to code in order to create SmartItems. I am happy to dispel this belief and assure to you they do not. While using code is one way to create a SmartItem, there are other methods that do not require it. Which method you use is entirely dependent on your capabilities and the type of SmartItem you would like to create.

**Figure 5.1.** Examples of Coded SmartItems, Drag-and-Drop Variable SmartItems, and Option-Based SmartItems




## Coding SmartItems

SmartItems can be written using code, which is utilized to accommodate and control SmartItem variables. In Chapter Two, I showed you a SmartItem for the Common Core State Standard for third grade mathematics. This SmartItem was written using code in a common language that is relatively easy to learn. (Learn more about this SmartItem here.)




## Drag-and-Drop Variables for SmartItems

Tools like the SmartItem API accomplish something similar without having to actually write code. The "coding" takes place behind the scenes, allowing item writers to simply create data sets and drag and drop their variables using a graphical user interface.




## Option-Based SmartItems

Some objectives are better measured by simply increasing the number of options, rather than producing code or using a GUI. The item designer/writer first decides on a selected-response item type, such as MC or DOMC (or any other format that requires the test taker to select an option). To create a SmartItem, the item writer can simply produce dozens, hundreds, or even thousands of options.

# Inside a SmartItem

## Correct Options

Let's take a moment and give you a window into the development of a SmartItem that has 324 unique renderings. Let's use the American History item from Chapter Two. For this SmartItem, the stem was invariant, with the text remaining the same in every rendering.

**"Is this statement true about how the amendments in the Constitution protect civil liberties and civil rights?"**

This SmartItem uses the "creating an abundance of options" method discussed above. When I created it, I wrote a single stem, then created 18 correct options. Each option correctly states how a single amendment protects civil liberties and civil rights. As a DOMC SmartItem, the test taker only sees one option at a time. This option is randomly picked from the many correct and incorrect options available. The examinee also sees YES and NO buttons, from which he or she will choose.

While the question has 18 correct options, it is easy enough to duplicate these options and simply reword each one. By changing the language of most or all of the options, you can easily increase the number of correct options from 18 to 36. *Doing this would frustrate individuals who might harvest, use, and sell your test content.* This principle holds true for incorrect options as well.

**"The Second Amendment protects the people's right to keep and bear arms"**

**DUPLICATED AS**

**"Keeping and bearing arms is protected by the Second Amendment."**

### PLEASE NOTE:

Duplication in this manner isn't necessary if the SmartItem truly covers an entire objective. If a SmartItem covers the breadth and depth of an objective completely, there is no need to take extra efforts to duplicate and increase the option pool. Let's exemplify this principle by using the American History SmartItem.

The objective for this item states an examinee must know all the civil liberties that protect civil rights. There are 18 amendments in the U.S. Constitution that protect civil liberties and civil rights. In this SmartItem, there is a correct option for each of those 18 amendments. Imagine one of these correct options is stolen and shared online. The information that was shared would not be useful to anyone. An examinee who tries to cheat using this information would be stymied and still wouldn't know which of the 18 amendments they would see on the exam; the stolen question would simply show information the examinee should

**PLEASE NOTE:**

already be studying. In the end, they will still have to study all of the 18 amendments that protect civil rights.

When a SmartItem is properly designed and created to cover an entire objective, it can't be stolen for any useful purpose. Under these circumstances, paraphrasing the correct options in order to create more correct options would not have any benefit.

## Incorrect Options

For SmartItems, an incorrect option functions in much the same way as "distractors" do for the traditional MC item. However, we refer to them by the more general term. Incorrect options usually come in two forms:

1. Re-wording the correct option to make it incorrect.
2. Writing another incorrect statement related to the objective.

For example, it is possible to create 306 incorrect options for the American History question by simply pairing any one of the 18 amendments with the protections stated in the other 17 amendments (18 x 17 = 306). For a question about the Second Amendment, you could substitute the protection afforded by the First Amendment.

**"The Second Amendment prohibits laws respecting an establishment of religion, and protects freedoms of religion, speech, and the press and the rights to assemble peaceably and to petition the government."**

While this SmartItem could be created using coding skills or a GUI, in this instance, it was created simply by manually increasing the option pool. By writing one stem, *18 correct options, and 306 incorrect options, you have created a SmartItem that can use this total of 324 options to render in a myriad of ways*. Security is further enhanced by using the DOMC item type. While coding or GUI technology might speed the process up, it is by no means necessary to achieve the desired results.

## Settings for DOMC Items

You will notice this SmartItem uses the DOMC format. You may be curious how, with so many correct and incorrect options, the settings are set and controlled. Luckily, SmartItem technology makes this easy. However many correct or incorrect options there are, the designer/writer can control their presentation and scoring through "settings." For Caveon's development and delivery platform, Scorpion, these settings are shown in Figure 5.2 below. (Please note: The red "S" followed by a number is simply a label to aid in this explanation.)

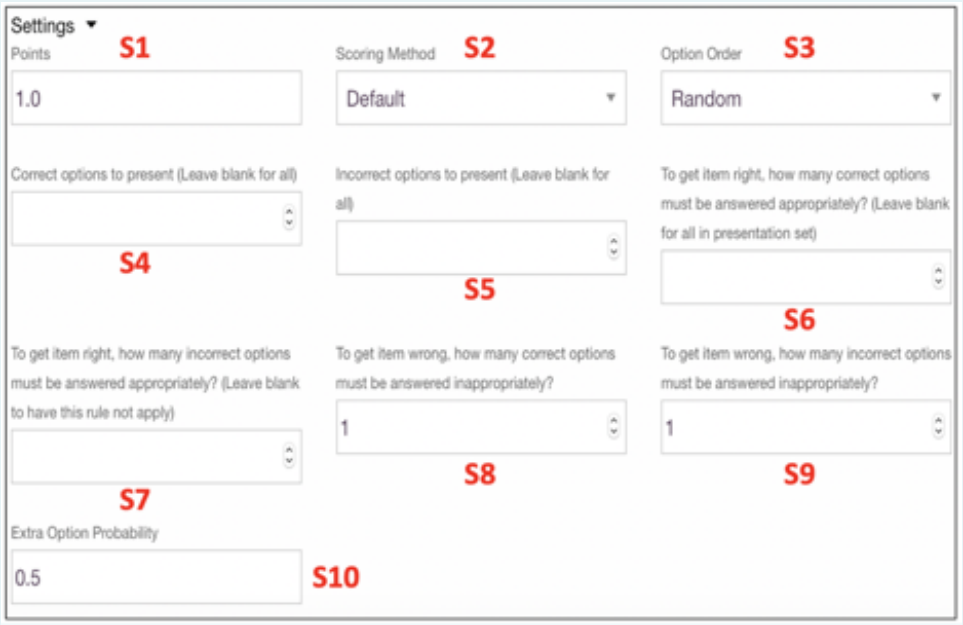



**Figure 5.2.** DOMC settings in Scorpion, Caveon's item development and delivery platform.

The top line of three settings (labeled S1, S2, and S3) shows default settings for scoring weight, scoring method, and option order. Each has a standard set of choices for weighting, scoring, and setting the order of options.

| Box | If Set To | Reason |
|---|---|---|
| S4 | 1 (do NOT leave blank) | The DOMC item will select 1 of the 18 correct options randomly. There may be times when you want to enter a number greater than 1. |
| S5 | 4 (do NOT leave blank) | The DOMC item will randomly select 4 of the 306 incorrect options. The number of incorrect options can also range widely. |
| S6 | 1 (or leave blank) | S6 is a scoring setting. In this case, I want to score the item as correct whenever the test taker sees a single correct option and responds with "yes." This number can also be greater than 1. |
| S7 | Leave blank | S7 is also a scoring setting. By leaving this blank, the item will not score itself or end based on answering each incorrect option with a "no" response. Keep in mind that it might be valuable to place a 3 or 4 in this box, meaning that a person could get full credit for the item without ever seeing a correct option, only incorrect options, which they answer correctly. |
| S8 | Almost always 1; set by default | This is a setting for scoring the item as incorrect. It is set as a default for 1 and should probably not be changed. It means simply that answering "no" to a correct option will result in the item being scored as incorrect. |
| S9 | Almost always 1; set by default | This is a setting for scoring the item as incorrect. It is set as a default for 1 and should probably not be changed. It means simply that answering "yes" to an incorrect option will result in the item being scored as incorrect. |
| S10 | Default is 0.5 | Unless you have a good reason, keep this setting at the default level of 0.5. It means that when the item ends and is scored, another random additional option (RAO) may be presented at 0.5 probability. |

**Figure 5.3.** Possible values for DOMC settings in Scorpion.

The remaining settings (S4-S10) require decisions the writer must make about any DOMC SmartItem. Figure 5.3 shows how I set the settings for the American history question, along with an explanation of why. This is a lot of information, but bear with me.

As you can see, given the settings and the number of options available, a specific correct option for the SmartItem is scheduled to be presented at an average rate of less[1] than one in 18 test takers. Incorrect options are scheduled to be exposed at the lower rate of less than one in 76.5 (306 options divided by 4) test takers.[2]

**This single DOMC SmartItem results in more than 1.2 trillion renderings or variations! If the MC format were used instead, the potential number of renderings would be slightly fewer, but still mind-boggling.**

## The SME Experience

Nearly every high-stakes testing program in the world utilizes subject matter experts (SMEs) to create their test items.[3] This is not an easy job; SMEs must combine their knowledge of the subject matter with their item-writing experience in order to create relevant questions that are both fair and can be scored properly.

In much the same way as traditional items, SMEs can be employed to create SmartItems. While expertise in the subject matter is always necessary, there is a key difference between producing SmartItems and producing traditional items:

### Designing vs. Writing

The most obvious difference between creating a SmartItem and creating a traditional item is that creating SmartItems is more about designing an item than it is about writing an item. For a SmartItem, SME's (working individually or as teams) start off by reading the skill description. They then envision how to design a single item to cover the entire skill. They brainstorm designs while carefully considering the requirements of the skill description. For example, for the third grade mathematics item from Chapter Two, the SMEs working together might decide the SmartItem must include:

1. Both multiplication and division problems.
2. An unknown value in any of the three terms of the equation.
3. Whole numbers (between 1-100). The SME (or SME team) must also decide how to render satisfactory incorrect options if a selected-response item type is used.

It is important to note SmartItems can be created for any item type. The main task for the SME is to decide how best to design the SmartItem so it completely covers the objective. First, the SME will decide which format—whether DOMC, MC, short answer, or drag-and-drop—the SmartItem will take on; the SmartItem could even launch a simulation to have the test taker complete a task.[4] In the end, each SmartItem will be

different and should consist of features that best target the desired skill.

Next, the SME will decide whether to include text, graphics, animation, video, audio, case studies, etc. These features have the potential to make the SmartItem more complex and can be used as needed to completely measure the skill on the exam. Writers should not be restricted by any perceived constraint that would not allow more than one item format to co-exist on the same test form—the power should be in their hands.



# 05 CONCLUSION

Accessing the appropriate technology needed for creating and utilizing SmartItems is relatively simple and inexpensive; programs interested in SmartItems can either switch to the robust software that supports SmartItems, or enhance their existing development and delivery systems by integrating with a SmartItem API.

SmartItems can be created in three different ways: coding, dragging and dropping variables, and creating an abundance of options. The SME selects which method is ultimately used based on their preferences and capabilities.

This chapter also illustrates the settings for properly formatting DOMC-based SmartItems. Additional item design considerations are outlined, including presentation and scoring rules.

## FOOTNOTES:

1. I say "less" because correct (and incorrect) options are not always displayed when a DOMC item is given to a test taker.
2. The likelihood that a correct or incorrect option is shown is less than these calculated odds because with DOMC items, not all of the options that are scheduled to be presented are actually presented.
3. Over the past 20 years, some items have been developed using AIG rather than SMEs. The economic value of AIG is to replace expensive human-based writing of items, creating larger pools of items for security purposes. Despite the promise of AIG, SMEs remain the primary method for developing items.
4. Many other item formats exist—see Sireci & Zenisky (2006).



Chapter

# 06 CREATING AND EVALUATING SMARTITEMS

## DESIGN, IMPLEMENT, AND EVALUATE

Chapter Five focused on the technology and procedures necessary for implementing SmartItems. This chapter takes a closer look at the processes for developing the content of SmartItems. Once you have the technology in place, how do you go about creating an actual SmartItem? How is the process different from developing a traditional item?

### Skill Descriptions[1]

The first step in exam development is to outline the descriptions of the specific skills that should be covered on the test. This task is usually accomplished by SMEs, who are experts in the skills or knowledge necessary for a particular job or instructional course. The SMEs list the most important skills or knowledge a qualified examinee should possess, and the result of this "skills analysis" is a list of objectives that should be included on the test. These skills will be weighted and ranked to establish the frequency with which they should appear on the test. (Some skills are more important, and should be given greater emphasis on the exam.)

When used for certification or licensure exams, this formal process of outlining skills for the test is often called a "job analysis" or "skills analysis." When the process is used for an education course, the terms that result are referred to as "learning objectives" or "competencies."

This skills analysis is a critical part of the test development process, and even more so when developing a SmartItem-based test. In traditional item development, skill descriptions are established quickly and are often written to be quite general. The vagueness of the objectives gives the item writers wiggle-room in



creating their items. There are potentially dozens or hundreds of items they could write for a specific objective, so the item writers have the ability to pick and choose what narrow range of topics they want to cover.

For SmartItems, however, creating skill descriptions is the **most important part of the development process**. In designing a SmartItem, the SME will need to create a question that covers an entire objective. As such, each objective needs to be well-thought-out and concisely stated. **Well-written skill descriptions lead to high-quality SmartItems**. Poorly-written and ambiguous skill descriptions make it almost impossible to create good SmartItems, or even good traditional items, for that matter. Here are examples of well-written and poorly-written skill descriptions:

## Good

The student will be able to recall the capital of every U.S. state.

This objective is clearly stated. It provides the skill ("recall") the child is expected to perform, and it indicates the expected level of mastery ("every U.S. state").

## Bad

The student will understand state capitals.

This skill description is vague and doesn't effectively describe what the student needs to do. It is also overly broad—specifically the term "understand." What does the student need to understand about state capitals? Which capitals should they understand?

## Worse

Capitals.

This description is only a single word. Not much for an SME to work from. It seems crazy, but in my experience, many a test has been written to objectives that are this sparse.

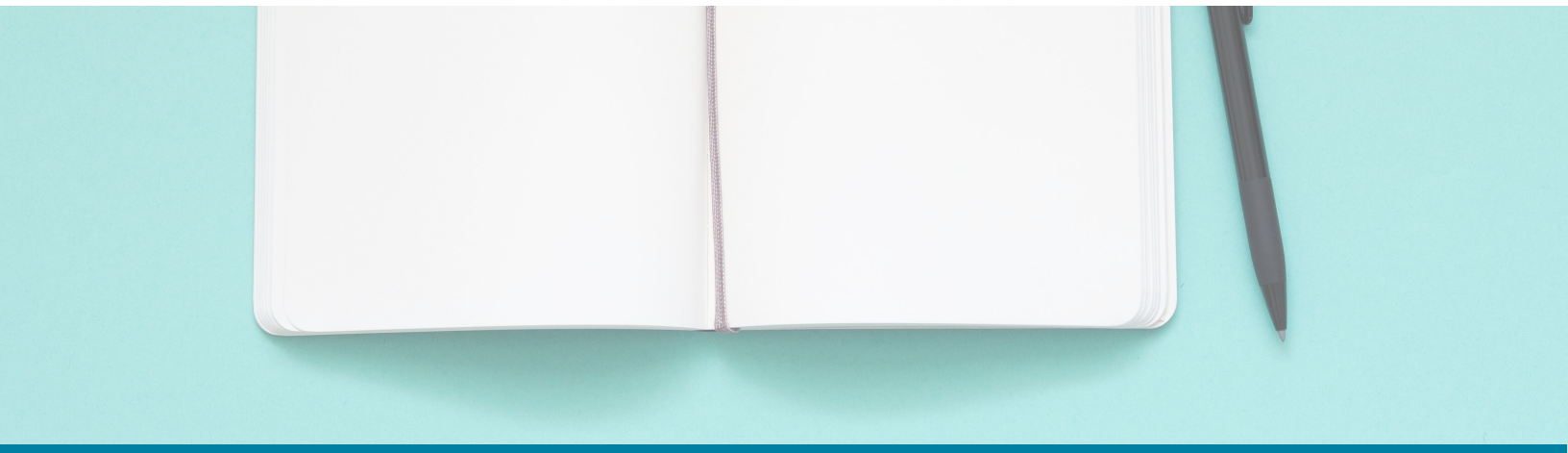
**WELL-WRITTEN SKILL DESCRIPTIONS LEAD TO HIGH-QUALITY SMARTITEMS.**

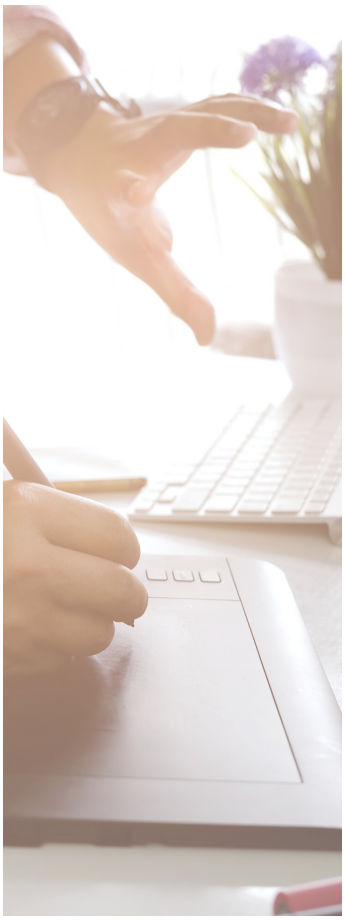
To produce great SmartItems, testing programs must take more time, care, and effort when outlining their objectives. Great objectives are specific. They outline the scope of the skills the test is intended to measure, and they offer guidance for the best ways to measure the skill.

## Designing Tests with SmartItems

For over 100 years, the most common type of test design has involved the creation of several "forms" for any given test. Different forms are made up of similar items, but with varying content, and questions are often re-ordered or rephrased. This is done to reduce item exposure rates and protect the exam's security. With this design, each test taker will randomly receive one of the forms. Each form typically has the same number of items, takes the same amount of time to complete, and is scored similarly. Minor differences between the forms are handled by statistically equating the forms and adjusting the scores accordingly. Some tests may have only a single form, others will have two to three, and others could potentially have dozens. The number of forms used is generally dependent on the number of test takers, the exposure rate of items, and other test security concerns. Since the widespread adoption of computers for creating and administering exams, new test designs have been used regularly. For example, Computerized Adaptive Tests (CATs) provide a unique experience for each test taker—the test actually changes during the exam, adapting itself to the abilities of each test taker. Linear-on-the-fly tests (LOFTs) create new forms for every test taker by selecting items from a larger pool of items. This process is randomized, either for the whole test or for a specific objective, ensuring each test taker sees a similar number of items for each content area while maintaining test security. This process is referred to more formally as "stratified randomization" (see Chapters Three and Four).

With SmartItems, as with CATs and LOFTs, multiple fixed forms are no longer needed. A single well-designed form—although NOT a fixed set of items—is sufficient. To illustrate, imagine a test made up of a single form of 70 SmartItems. These 70 SmartItems cover 70 distinct skills. Because each SmartItem renders in thousands of ways, each test taker (whether there are 10 or 10,000 test takers) will see a unique set of questions, even though they see the same form. As with CATs and LOFTs, the scores of each test taker are comparable if the test has enough items, and those items are presented according to the test design.





SmartItems are not test designs from CATs and LOFTs. Rather, they can be used with these and any other test designs. A CAT continually estimates a test taker's ability level and then selects the next item the test taker will see from a pool of items of the appropriate difficulty level. SmartItems can be used in tandem with any CAT test. Keep in mind the distinguishing feature of SmartItems is that they are true items. The SmartItem itself, rather than its renderings, has statistical properties needed by the CAT. Because skills can be easy or difficult, and SmartItems are created to be congruent with those skills, the SmartItems themselves will have the same range of average difficulty as the skill. This is calculated using the usual item statistics. During a CAT, when an estimate of the test taker's ability is approximated, a SmartItem matching that ability level is selected and presented.

The number of SmartItems needed for a CAT is much smaller than the traditional number of items. The item pool for a CAT can simply contain a single SmartItem per each skill the test covers. For any ability estimate, content balancing algorithms can still be used to make sure every CAT covers the content in the same way. Regardless of test design (fixed-length form, CAT, or LOFT), using SmartItems significantly reduces the total number of items needed for the test.

## Designing SmartItems

It should be clear by now that creating SmartItems is much more a process of "designing" than of traditional item "writing." SMEs must design a SmartItem to cover the complete objective provided to them by the original skill analysis. Let's use a randomly selected K-12 state standard to exemplify this process. This standard or objective below is found here, or at http://ritter.tea.state.tx.us/rules/tac/chapter115/ch115b.html.

> The objective states:
>
> "The student is expected to identify ways to prevent substance misuse, including the misuse of prescription and over-the-counter drugs, and substance use disorders."

Imagine you're a health sciences SME. Your task is to create a SmartItem that covers this entire skill. Your first step would be to brainstorm what a SmartItem for this skill might look like and what factors it needs to cover. Below, I've listed out my personal thought process for approaching this SmartItem. Other SMEs might approach the design task differently. There is no right way, so long as the content ends up covering the entire skill!

## A WINDOW INTO MY THOUGHT PROCESS:

**A** I think a selected-response format would work well since the skill is about "identifying ways." In order to avoid testwiseness associated with a multiple-choice question, I'm going to use the DOMC format. Other formats such as matching might work as well. I'll keep that in mind.

**B** The skill states the student must identify ways to prevent BOTH (1) substance misuse and (2) substance use disorders. Therefore, I'd better make sure my SmartItem allows a student to identify ways to prevent both of them. For now, I'm going to divide this into two separate areas and outline the content separately for both substance misuse and substance use disorders.

**C** I'm going to start with the first component—ways to prevent substance misuse. To start, I'll create a list of commonly misused drugs (illegal drugs, prescription drugs, over-the-counter drugs, etc.), then outline a list of common misuses for each of these drugs. I will be as thorough as possible.

**D** Because this is a DOMC-based item, the SmartItem will need to present incorrect options. I'll now write a list of incorrect options. This list will include incorrect or ineffective ways to prevent substance misuse. I will be as thorough as possible.

Okay, I've written my content for "preventing substance misuse." I'll now follow the same process for "preventing substance use disorders."

**E** I now need to combine both of these skills into the same SmartItem. What proportion of SmartItem renderings should ask about preventing substance misuse compared to preventing substance abuse disorders? The skill description doesn't provide a lot of guidance, so I'll design the SmartItem to render them equally.

**F** I've now outlined my content and have a design plan. Time to start creating.

The above list was a window into my thought process for this SmartItem. I am the first to admit I may have missed design considerations, and there may have been a better route I could have taken. However, I think I did a decent job outlining and covering the necessary components of the content provided in the skill description.

Please note that creating SmartItems is easier if they are planned for from the very first step of the development process. This allows the skill descriptions to be thoroughly thought out and written down to include all the information necessary for creating a great SmartItem, including the components of the skill that should (and should not) be covered, sample designs of the item, and perhaps the target difficulty level the writer should aim for. Attaching this information to the skill description not only expedites the SmartItem creation process, but it also helps the SMEs designing the objectives to have a clearer and more comprehensive vision for the overall content of the exam.

Now on to the next step—creating the "scaffolding" or "road map" of the SmartItem rendering, including the variables and option sets.



# An example of what the SmartItem might look like:

Variable designed to change each time the SmartItem is administered.

## STEM:

Is this one strategy for preventing the misuse of _____?

## PARAMETERS:

1. (25% chance of selection) Misuse of < OTC drugs list, including cough syrup, loperamide, and others >
2. (25% chance of selection) Misuse of < prescription drugs list, opioids, benzodiazepine, etc. >
3. (50% chance of selection) The substance use disorder < substance abuse disorders list >

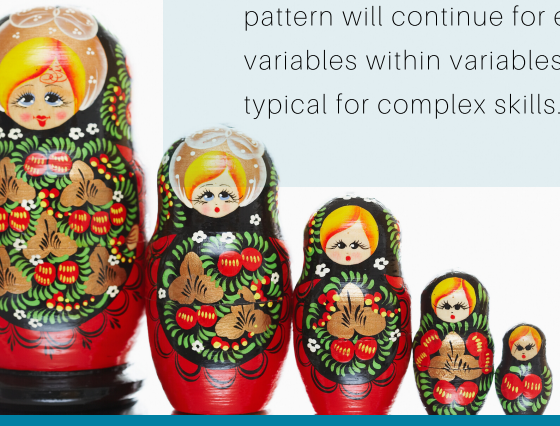
If "misuse of cough syrup," for example, is selected to appear in the stem, a possible option set could include a list of correct prevention alternatives:

A. (First correct preventative measure)
B. (Second correct preventative measure)
…
N. (nth correct preventative measure)

If "misuse of loperamide," for example, is selected to appear in the stem, the possible option set of correct prevention alternatives could include:

A. (First correct preventative measure)
B. (Second correct preventative measure)
…
N. (nth correct preventative measure)

Incorrect options for either stem can be pulled from the correct options from the other stem or from a separate list of incorrect options specific to that stem. This general pattern will continue for each variable that appears in the item. In this item, we have variables within variables, like Russian dolls. Having embedded variables is typical for complex skills.



This example is just one of the many approaches that could be utilized for this SmartItem. Another SME might come up with an entirely different way to design a question that accomplishes the same goal.

Clearly, the process of designing SmartItems is very different from designing traditional items. Traditional items require minimal—if any—design work. Perhaps the SME will need to choose the item format (although an overwhelming number of items continue to use the ubiquitous MC item format with a single correct option). In addition, the SME will need to decide which specific content to include. For the skill provided above, a writer might decide the traditional item will measure the student's ability to identify a correct way to prevent over-the-counter substance misuse of a single over-the-counter drug: cough syrup. Perhaps they choose cough syrup because they have recently read research on the misuse of this particular OTC drug. They choose a correct answer, and it is presented in a selected-response-based item along with three to four incorrect options. The item is then considered complete.

For traditional items, the SME's main goal is to look at the objective and then arbitrarily decide which sliver of information the item will cover. With SmartItems, there is a complete paradigm shift. The main goal of this design phase is to organize multiple pieces of information in the most logical way possible. Editors, reviewers, managers, and coders should be able to understand how each variable is connected to the stem, to the correct and incorrect options, and to the other variables in the item. While there are no concrete rules for this process, the overarching guiding principles are:

### Creativity

SMEs should have the freedom to design and create the best SmartItem for the skill.

### Coverage

SMEs should focus on covering the entire objective.

### Clarity

SMEs should organize their information in a way that is clear for editors, managers, and coders—and ultimately, for test takers.

The use of data from tables, charts, and spreadsheets can help organize complex items. Keep in mind that creating SmartItems is a new test development skill. Over time, we will refine this process and establish best practices.



## Building the SmartItem

Once the SmartItem is designed, technology is required to build it. The technology tools[2] must

1. Allow code to be produced within item components,
2. Have an easy-to-use graphical user interface (GUI) and/or
3. Include a way to expand the number of options manually.

The code used in Caveon's tools is well-known and relatively easy to learn, but most SMEs will not have the time nor inclination to learn it. The case studies described in Chapter Seven describe how SmartItems can be created with a team of non-coding SMEs, supplemented by a coder to help with implementing the design. This is one method for creating SmartItems. Other methods, such as using a GUI or expanding the number of options, do not require coding at all.

As with traditional items, SMEs are also responsible for creating the scoring rules for SmartItems. These scoring "keys" are a necessary part of whatever item format (e.g., DOMC, MC, etc.) was selected for the SmartItem.

Once created, the SmartItem must be checked and approved. This means technology must be in place to view multiple renderings of the SmartItem and confirm its performance. The renderings should be presented one at a time, should allow the SME to select correct and incorrect responses, and should produce the correct score. In Caveon's tools, this is referred to as "previewing." Its purpose is to show the



item developers that the design is working.

Two roadblocks stand in the way of successful SmartItem generation:

1. Creating code. SMEs cannot be expected to learn to code, and most programs will not have access to coders that can help. How then can a testing program write these types of SmartItems?
2. Test Administration Limitations. Once developed, a testing program might not be able to use their SmartItems and because of test administration limitations.

To remove these roadblocks, Caveon has created a SmartItem API. This application can be integrated with any program's current delivery and development systems. It provides

To remove these roadblocks, Caveon has created a SmartItem API. This application includes a GUI interface that allows SMEs to create SmartItems without using code, and it can be integrated with any program's current delivery and development systems. It provides a testing program to develop SmartItems within their current test development tools, and to administer those SmartItems through their current administration system. The SmartItem API was built to make the benefits of SmartItems available to programs that wish to avoid changing their testing infrastructure.

## The SmartItem API and Its Graphical User Interface

In 2019, Caveon released a tool that makes it possible to create and use SmartItems on any development and test administration engine. This tool is called the SmartItem API. Generally, API stands for Application Programming Interface, and allows applications to communicate with one another. The SmartItem API makes it easy for any program (or its vendor) to connect its development and delivery systems to interface with it. Once connected, it allows SmartItems to be created and administered within the program's environment, using the program's preferred variety of item types.

In Figure 6.0, you can see the SmartItem API and the drop-down menu displaying the available item types. This screen is accessed through any test development software. For this example, the API is connected to Caveon's test development and delivery platform Scorpion, and all the item types supported by Scorpion are shown in the SmartItem API menu. The list of item types may be longer or shorter if the SmartItem API is connected to another system.

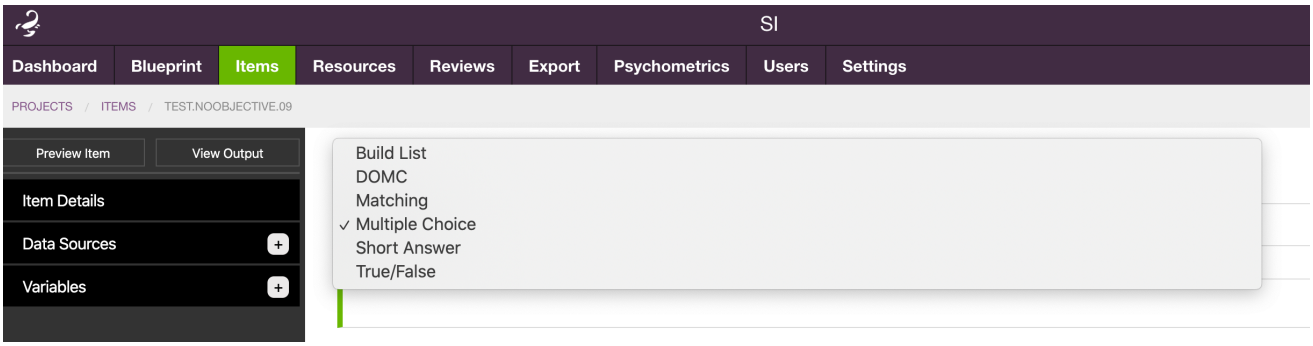



**Figure 6.0**
SmartItem API screen displaying the drop-down menu of available item type templates.

## The general steps for using the SmartItem API:

# STEP 1

### Choose an Item Type (or Template):

Once connected to the SmartItem API user interface, the SME chooses which item type will best outline the skill being measured. The available options are contingent on the item types that are supported by the program's development tools. (See figure 6.0.)

# STEP 2

### Design the SmartItem:

Using the SmartItem API, the SME designs a SmartItem using this list of tools:

### A Data Sources

Data sources are spreadsheets that contain lists of data. The SME creates these data lists based on content that needs to be included in their SmartItem. The data source could be a list of people or things, or it could be full phrases worded to be correct or incorrect. The data source could also be images, audio files, or video files. An SME can create and use more than one data source for a SmartItem. Figure 6.1 is an example of a data source used for a question that asks test takers about the location and name of state capitals.

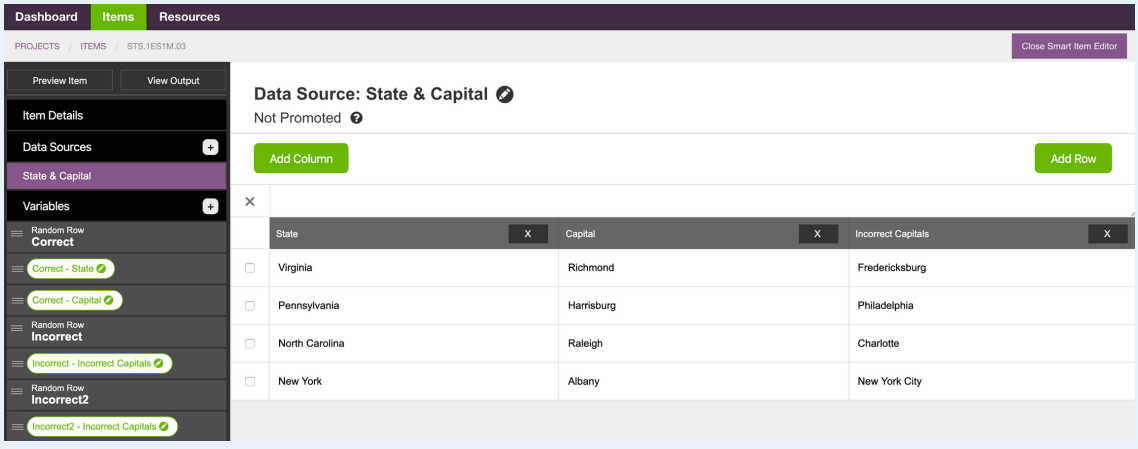



**Figure 6.1**
A data source showing data for the capitals of every U.S. State. Only a few states, listed in reverse alphabetical order, are shown.

# The general steps for using the SmartItem API:

B

## Variables

The SME then selects and names different types of variables. These variables make it so information in the data sources will be used in the SmartItem. However, a variable could also use mathematical formulas or a variety of other functions. Figure 6.2 shows a list of several variables to support the States/Capitals SmartItem. It also shows how to add a new variable.

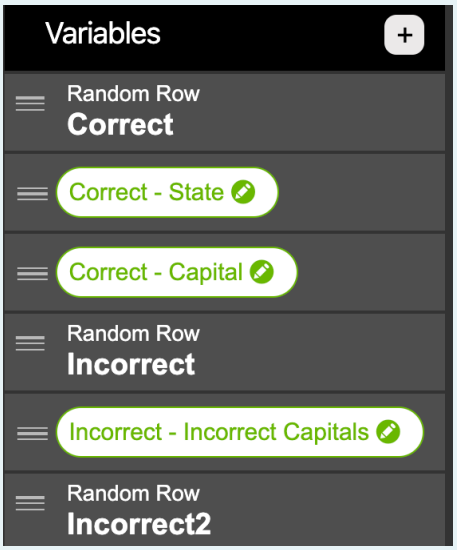



**Figure 6.2**
A list of SmartItem variables for a particular SmartItem. New variables can be easily added.

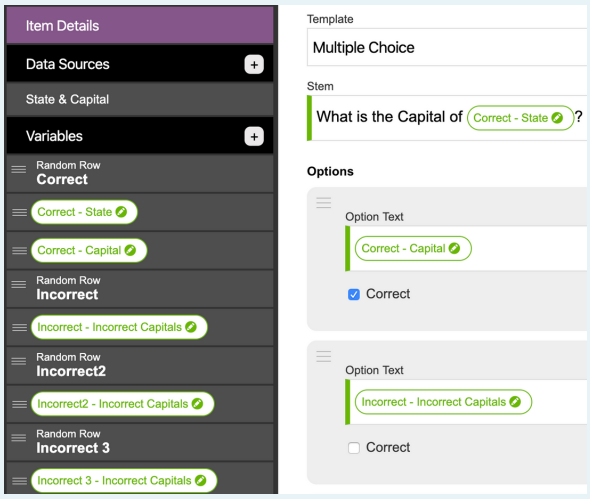


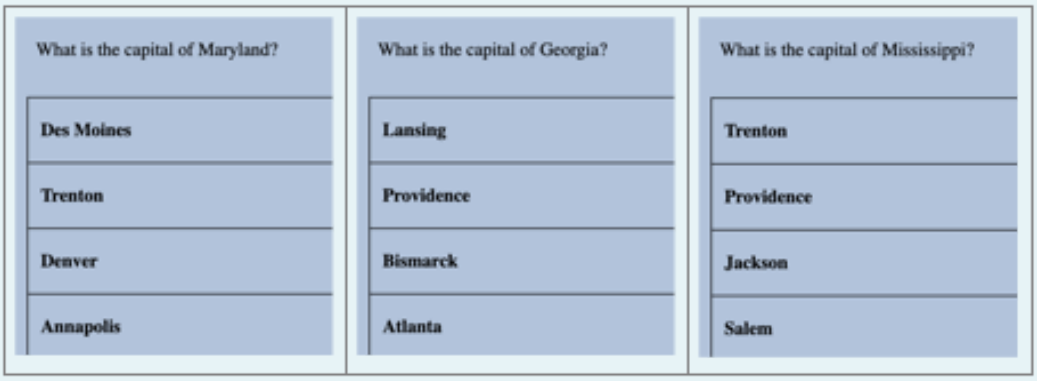

**Figure 6.3**
A list of SmartItem variables for a particular SmartItem, and illustrations how they are used in the MC item stem and options.

Figure 6.3 shows the stem and three out of five options for the question. It shows how the variables are dragged into the stem and the options of the SmartItem. These variables will now randomly draw information from the data sources every time the SmartItem is rendered.

With the ability to combine variables and data sources, an SME with no coding inclination or experience can design and create virtually any complexity of SmartItem. Figure 6.4 shows three of the hundreds of random renderings for the SmartItem above.

## The general steps for using the SmartItem API:

**Figure 6.4**
Three different renderings of the States/Capitals SmartItem. For this figure, these renderings are shown in QTI display format (QTI refers to the International Standards for Question and Test Interoperability.




By using the SmartItem API, any test program can begin to use SmartItems in their high-stakes operational exams, in low-stakes practice exams, or for research purposes. For some, the SmartItem may need to be field-tested or reviewed for accuracy.

Please note that describing the dynamic process of SmartItem creation through static screenshots leaves much to be desired. To see the process in action, please visit Caveon.com/Smart.

## Expert Reviews, Psychometric Reviews, and Field Testing

Because SmartItems render randomly in so many ways (there could be literally millions of renderings for a single well-designed SmartItem), they present a challenge for item reviewers. How can a reviewer evaluate a SmartItem if it changes each time it is viewed? Should they review thousands or millions of renderings? Such a review would be infeasible and impractical.

Evaluating individual renderings is not the best way to review a SmartItem. Rather, the reviewer should analyze a SmartItem to make sure it properly measures the identified skill. A logical place for this review to occur is during the design and construction phases of the SmartItem development process. The reviewer would look at the item scaffolding, the variables, the data sources, and other SmartItem metadata to ensure they reflect the design intended by the SME. A careful review of this nature provides evidence of validity.

Following this initial step, the reviewer should ensure a SmartItem produces renderings that accurately represent the skill being measured. They should also ensure there are no bugs in

the settings or code that cause the SmartItem to malfunction. This is possible using the preview function in the SmartItem API. Such a review can take a look at any number of random renderings until the sample size is large enough to support the accuracy of the SmartItem.

Editors looking for smaller textual faults in the SmartItem can then read the portions of the SmartItem that contain text and/or images to make sure they are free from errors. Indeed, comments during these reviews generally fall into the following four categories: editing issues; coding and functionality issues; inaccuracy, congruence, relevance, or difficulty issues (a.k.a., CARD); or no issues at all. Figure 6.5 is an example from a recent SmartItem development project of how these comments broke down:

In my experience, SMEs have generally enjoyed this review phase, especially those who participated in the development of the SmartItems. It is rewarding and somewhat astounding to see the concept of a SmartItem become a reality and to watch as the computer produces one rendering after another. Likewise, as the graph above indicates, seeing their item from an examinee's perspective helps SMEs further refine the content.

While this review process functions differently than the review process for traditional items, it has multiple benefits, and it is key when performing quality assurance checks. When the reviewers have completed their various evaluations and the SmartItem has been revised, it can then be subjected to a field test.

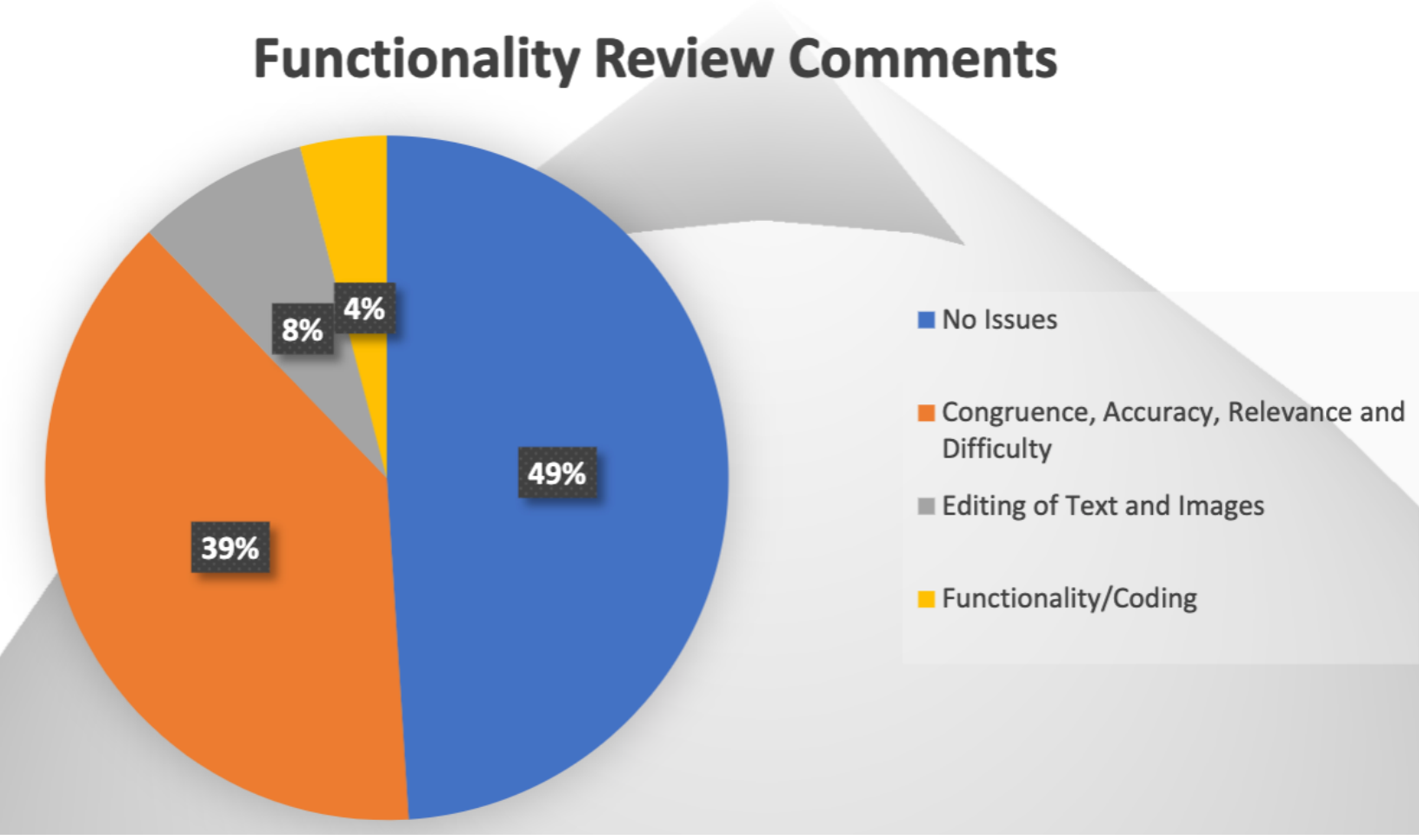



**Figure 6.5**
SmartItem review comment breakdown.




## Collecting and Analyzing Response Data from Field and Operational Tests:

The next step of the SmartItem development process is to test the SmartItem operationally. In administering a trial test, it is possible to record each test taker's answer for one or more SmartItem renderings.

This information can be helpful if legal steps require a complete reproduction of the testing event. Such a perspective of the test should help those in the legal system determine if the score provided is appropriate for the test taker's performance level and if the test was administered unfairly.

There is one benefit of SmartItems that has not yet been mentioned elsewhere in this book: An individual's test can be reviewed in complete detail by any audience without exposing the item and compromising future test results. You can give as many field tests as you would like, and you would not over-expose the SmartItem. It would still be valid for official testing.

However, the primary reason for conducting a field test is to conduct statistical analyses to find out if the SmartItems are performing to established psychometric standards. With the appropriate data, IRT or classical statistical analyses can be run on SmartItems. This process does not depend on knowing which specific rendering of the SmartItem the test taker saw—the data will be used to evaluate the SmartItem holistically.

To run these psychometric analyses, the test administration system must be able to display renderings, collect responses, and also record and store responses to SmartItem renderings. The data that should be collected (in real time) and stored include:

- the test taker's responses (choices, words typed, objects moved, etc.)
- the time it took for the test taker to respond (often called latencies or response times)
- the order of appearance of SmartItem components, whether simultaneous or sequential
- the use or view of SmartItem and test components (breaks, images, language translations, etc.)

In addition to operational testing, field testing is an important way to evaluate the performance of SmartItems empirically. These field tests are sometimes referred to as "tryouts," "beta tests," or "pre-testing." The data from these field tests are then evaluated to determine whether the item is too difficult or too easy, whether it discriminates between test takers of different abilities, whether it is fair, whether it takes too long to answer, and whether it is easy to guess.

Caveon's exam development and delivery tool, Scorpion, is capable of presenting SmartItems to test takers for field tests. Scorpion also collects and stores the data necessary to perform the appropriate statistical analyses, and even automatically provides a real-time version of those analyses. However, the SmartItem API can help a program that doesn't use Scorpion to create and administer field tests without having to replace its technology infrastructure. The SmartItem API does not, however, collect data. Data are contingent on the delivery tool being used. It is possible a program's systems cannot collect, store, or retrieve the amount and variety of data necessary to support SmartItems. When this is the case, Scorpion can serve as an exceptional resolution.



## Creating SmartItems:

### One SME's Experience with SailPoint Item Development Workshops

SailPoint is an information technology company specializing in identity governance, a technology that plays a crucial role in modern enterprise security. Recently, they decided to use SmartItems for their certification exams. In the "Excellence Edition" of *The Lockbox* e-magazine, Caveon's own Chief Editor, Tara Williams, wrote an article about her experiences creating SmartItems for these exams. You can read the entire article yourself, but here are selected highlights of the relevant points Tara made:

It's been challenging, rewarding, and eye-opening—a process of perpetual discovery. Personally, I've taught subject matter experts how to write them, editors how to review them, and have recently been thinking about the ways SmartItems will influence—and possibly improve—our test development processes.

Based on my experience, I believe that as SmartItems are adopted, programs will discover that, beyond the core benefits of reducing test theft and cheating threats, SmartItems have fringe benefits—more specifically, quality control during the development process.

### SmartItems demand clear learning objectives

In our SmartItem writing workshop, we ensured that at least one expert who created the objective domain attended. We witnessed spirited, important conversations between the objective domain creator and the writers about the scope of the exam, about what we were trying to measure in objectives, and why. Observing these conversations was heartening, as we were witnessing further refinement of the objective domain. We were witnessing quality control of the exam, demanded by the SmartItem format.

### SmartItems limit writer subjectivity

In our workshops, our writers felt empowered as they worked back and forth with the creators of the objective domain to ensure the content in their SmartItem encapsulated the entire objective. Additionally, being passionate about their area of expertise, they liked the idea that candidates would have to study the material more deeply and thoroughly since exam items cover all content associated with the objectives.



**Tara Williams**
Chief Editor, Caveon
Test Security





## AUTHOR'S NOTE:

### Personal Observations from the SailPoint SmartItem Writing Experience

How did this experience differ from the typical workshops we've done, and how did the SMEs take to the new tasks? (Don't tell anyone, but we were not sure ourselves what some of those tasks would be when we got started!) Most of the plans of the workshop were based on logic, not experience. We all had our fingers tightly crossed.

I don't want to keep you in suspense, so I'll tell you now that the workshop was a great success. It was a pleasant and fulfilling experience for me, and—as far as I could tell—for everyone else involved. The SMEs accepted the task of creating "items" that measured each objective as completely as possible. They focused their attention on the skills to be measured, not on the format and rules of multiple-choice items. We were mentors for them at all times, there to teach them what is possible, and to support them in their vision.

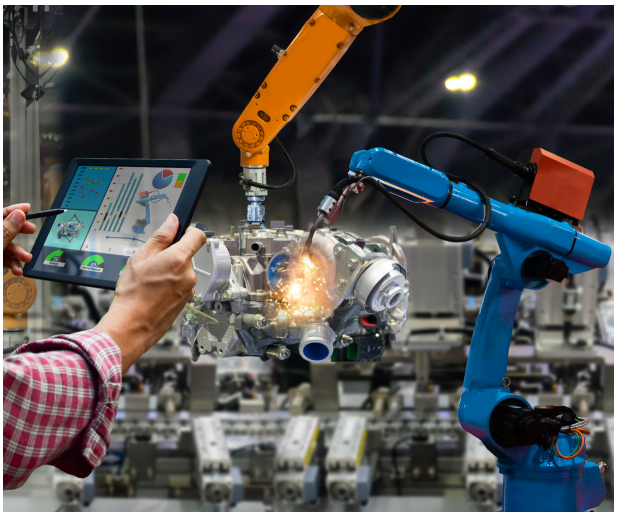
While it seemed to me that we produced a stellar set of items, it is exciting to know we can continue to improve. While this was the first-ever SmartItem workshop, for all its success, we learned much about how the process can be improved.

SailPoint was supportive and assumed the risk of innovation with as much enthusiasm as we did. The SMEs were engaged for three days, and each one expressed how fun it was to participate. They realized we were trying new things, accepted the challenge, and trusted us to lead them.

One of the most surprising things I learned was SmartItems are not as difficult nor as time-consuming to create as one would think. In fact, if a development project calls for more than one traditional item for a skill, it might be easier to create a SmartItem for the skill. If SME reports from experienced SMEs are to be believed, the exercise can even be fun.

## Reviewing SmartItems

Traditionally, items are reviewed by SMEs to confirm accuracy, find potential bias, and edit language. When conducting the workshops, we needed to make adjustments to the usual process; it is not possible for a reviewer to "read" and evaluate a SmartItem like they would a traditional item—a SmartItem renders in too many ways, making it impossible for a reviewer to "read" more than a small number of them.



Rather than follow the traditional item review process, we created a new system similar to the quality assurance process manufacturers use to evaluate products. Rather than evaluate every product, they evaluate a "sample" of the products. In our case, the reviewers first judged the quality of the code or design of the SmartItem. They then "ran" the item as many times as necessary, reviewing a sample of random renderings until they were comfortable the SmartItem was working as designed.

**RATHER THAN FOLLOW THE TRADITIONAL ITEM REVIEW PROCESS, WE CREATED A NEW SYSTEM SIMILAR TO THE QUALITY ASSURANCE PROCESS THAT MANUFACTURERS USE TO EVALUATE PRODUCTS.**

By reviewing the "inner workings" of the SmartItem, the reviewer becomes more confident that potential bias has been eliminated, the item will be fair for different populations, and significant errors have been caught and fixed. As with traditional exams, the utmost care was taken to ensure we produced quality SmartItems.

## Setting Cut Scores for Tests with SmartItems

Not being able to review items in the traditional way also poses a challenge for the popular Angoff (or Modified Angoff) standard-setting procedure. The Angoff method requires that "judges" read the item and then evaluate its difficulty, usually by estimating the proportion of qualified individuals who would answer the item correctly. From those judgments, a pass/fail score (called a cutscore) is calculated. Since SmartItems constantly render unique views of a SmartItem, it is impossible for an Angoff judge to make consistent judgments.

There are at least two useful methods that can be used to set cutscores for a test of SmartItems. One of these is a commonly used alternative to the Angoff, the other is a more extensive modification of the Angoff procedure:

# Methods for Setting SmartItem Cutscores

## Method 1:

### Contrasting Groups Cutscore Setting Procedure

The SailPoint exams are certification exams, and therefore require a pass/fail score. To come up with this cutscore, we devised an alternative to the Angoff called the Contrasting Groups method. This alternative did not rely on the judgments of reviewers, but instead on the actual test performance of certification candidates. In a nutshell, the Contrasting Groups method assembles two groups of test takers, one judged to be "qualified" and the other judged to be "unqualified." Individuals from both groups take the test comprised of SmartItems. Their test scores are compared by creating score distributions for each group on the same score scale.

Figure 6.6 shows what the results of this procedure would typically look like. As you can see, the two groups overlap. This is not an uncommon result. The cutscore is set by selecting a value between the two groups that maximizes the accuracy of certification decisions. For this data set, setting the pass/fail point at a score of around 12 would seem reasonable. That dividing point is illustrated with the red vertical line.

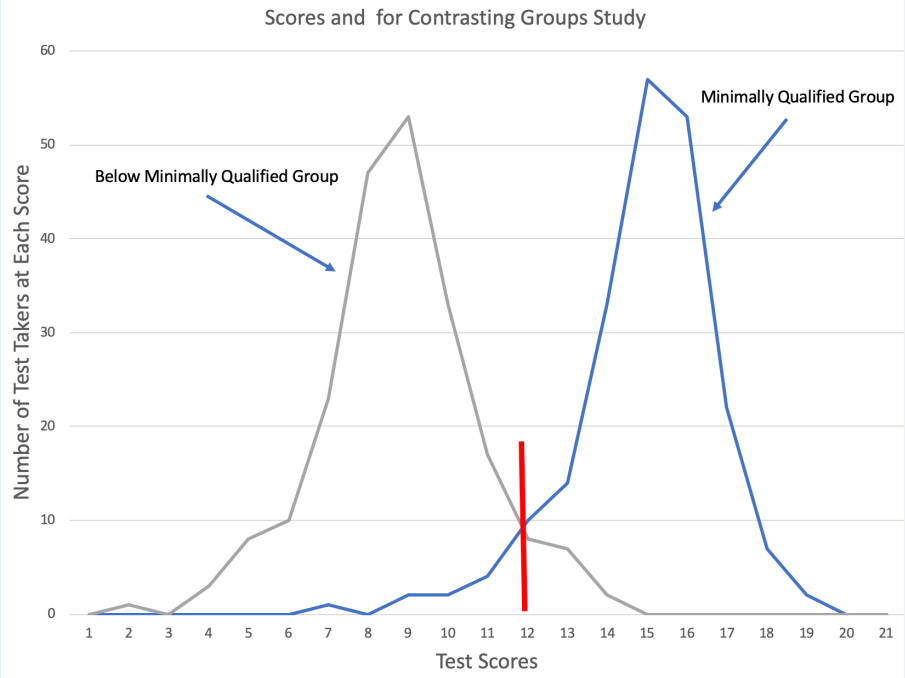



**Figure 6.6**

Results from sample Contrasting Groups standard setting procedure.

# Setting Cut Scores for Tests with SmartItems

The Contrasting Group method does not use the subjective opinions of judges, but instead relies on empirical results from actually taking the test. It is a reasonable and effective method to use for setting cutscores on SmartItem tests. To work, the Contrasting Groups method requires items that can be scored right or wrong and the ability to create a test score. A test with SmartItems meets both of those criteria, as do almost all other tests.

## Method 2:

## IIT-Enhanced Angoff Judgments

For one SailPoint exam, the Contrasting Groups method could not be used because not enough individuals could be recruited to take the test. This particular exam required highly technical skills, which only a few individuals possessed at the time the exam was developed. Instead, the Angoff method was adapted and improved. First, the judges evaluated the objectives rather than the renderings of SmartItems. This change seemed reasonable since the SmartItems were congruent with the objective. By evaluating the difficulty of the objective, the judges were essentially evaluating the overall difficulty of the SmartItem.

After the judges evaluated the objectives for the exam, Information Integration Theory (IIT) was applied to the judgement process. IIT is a cognitive science theory specifically designed to confirm the accuracy of subjective human judgments. It does this by introducing judges to an integrated cognitive judgment task. For the SailPoint exam, the two integrated factors are the text of the objectives and the proficiency levels of certification candidates.There were 65 objectives, and two proficiency levels: Minimally Qualified Candidate (MQC) and Below-MQC.

To use IIT for setting cutscores, the text of one objective is randomly presented to the judge, along with a specific proficiency level. The judge then estimates the difficulty on a scale of 1-100 for the assigned proficiency level. The judge is then be randomly assigned another objective and proficiency level. The task continues for all randomly-presented pairings of objectives and proficiency levels. In total, each judge would provide 130 judgments (65 objectives x 2 integrated factors).

## Setting Cut Scores for Tests with SmartItems

The findings from each judge, and from all judges combined, are then analyzed using a two-way ANOVA. The results of this confirm the consistency of individual judges and the overall evaluations of the group. With those results in hand, a cutscore separating the proficiency levels can be easily determined.

In general, this IIT-enhanced procedure avoids many of the criticisms of the various versions of the standard Angoff procedure.

- It is quicker.
- It avoids the back-and-forth discussions between panel members.
- It acknowledges that the differences between judgments are important.
- It statistically confirms the performance of each individual judge.

For the SailPoint exam, these advantages allowed a cutscore to be confidently set.



# 06 CONCLUSION

Utilizing SmartItems and realizing their vast array of benefits will require a few unique steps and a strong foundation of new technology. But the further steps needed for creating and reviewing SmartItems and setting cut scores can easily be accommodated by SMEs, and the technology needed for implementing SmartItems is painless to acquire and integrate, and fairly simple to use.

More than anything, the greatest barrier to successfully implementing SmartItems is a mental one—many programs will be unwilling to move away from the traditional methods for item development that have been used for decades. But these methods can (and should) be abandoned when more capable procedures and innovations are introduced. If testwiseness, answer-copying, and pre-knowledge could no longer negatively impact test scores, why wouldn't we adapt?

## FOOTNOTES

1. Recall that for this book, the generic term "skill" is being used as a synonym or substitute for these other terms: objective, competency, proficiency, construct, ability, behavior, performance, performance standard, and several others. While there are differences between these, they can all be represented by the generic term, "skill."
2. Because SmartItems are new, tools available to create and administer SmartItems are rare. Because the capability doesn't exist in most test administration and development systems, Caveon created Scorpion to design and build SmartItems and SEI to present them in the context of exams. Plus, Caveon created the SmartItem API, which easily integrates with other third-party systems to allow SmartItems to work on any other technology, without the need to replace what a program is currently using.

Chapter

# 07 EVIDENCE FROM CASE STUDIES

## EARLY ADOPTERS AND THE SMARTITEM IMPACT

Preceding chapters made the case that SmartItems should work. Randomization was shown to be an effective control method to equalize differences between items, including the random differences in difficulty introduced into a test by the SmartItems themselves. According to modern test theory, those differences balance out "over the long run," which means they balance out during the course of an exam. And really, using randomization for SmartItems isn't much different from what we currently do when creating equivalent forms or pulling items from a pool for CATs. It's the next logical step.

We have also made a case for why we should use SmartItems. SmartItems solve several critical validity issues, including testwiseness, effective cheating, and the theft of test content. Inviting a bit more randomness into our exams seems like a fair trade-off for solving these problems.

However, you probably still have dozens of questions about SmartItems in practice. These might include:

- How do SmartItems perform?
- What are their item statistics?
- Do they contribute to the reliability of a test?
- Do they add to the validity of a test? (That is, do they improve the ability to make valid inferences from test scores?)
- Have they been tried out on students or certification candidates?
- Have they been tried out on actual test takers in high-stakes exams?
- What evidence do you have for the claims?

The next few chapters will endeavor to answer these questions (and more) with evidence in the form of case studies, scientific research findings, and simulations.

## What works:

A quick Google search reveals a number of "What Works?" websites. One is titled, "What Works Clearinghouse." Its goal is to provide educators with the information they need to make "evidence-based decisions." To be "evidence-based" is to provide evidence that indicates what you are claiming actually works. There is no set standard on how much evidence is enough, and for new innovations, not as much evidence should be expected. Instead, the evidence should be compelling and interesting enough to encourage a further look.

SmartItems are still new. While some research has been done, more will be needed. SmartItems solve important problems but also raise many new questions about the way "things have always been done." Academicians, students, and practitioners from all areas of testing must join the effort to learn more about this innovation. New ideas and new evidence are always welcomed.

This chapter contains descriptions of two case studies. Each covers a different certification exam that was developed and published by the information technology company, SailPoint.

It would be helpful at this point to describe the difference between a case study and a scientific experiment. A case study is a set of conditions that are observed rather than controlled. As a result, a case study cannot determine cause and effect or rule out alternative explanations for what is observed. On the other hand, an experiment uses procedures to control for extraneous influences so the actual cause and effect can be seen and confidently concluded (at least within a small degree of acceptable statistical error).

While experiments are generally preferred, they are not always possible. One benefit of case studies is they allow us to study operational exams in a real-world setting. This is what we will see in the two case studies provided in this chapter.

## The Case Studies

The two case studies in this chapter describe the development and implementation of certification exams for the information technology company, SailPoint. SailPoint specializes in identity governance security, and because of that, is a natural advocate of test security. As such, it wanted to avoid the test security problems plaguing other IT certification programs. Specifically, SailPoint wished to prevent the rampant theft and online publication of test questions that often occur within days or weeks of a certification exam publication. An entire "braindump" industry has been spawned and supported by this type of easy, low-risk test fraud. As a result, many unqualified examinees are able to cheat on IT certification exams using information they found on the Internet.



Hoping to avoid this scenario, SailPoint welcomed the opportunity to try out SmartItem technology. The first exam (Case Study #1) was published in June of 2018 and is called the IdentityIQ Engineer Exam. The second exam (Case Study #2) was published in December of 2018 and is called the IdentityIQ Architect Exam. Both certification exams used SmartItems exclusively.

# 1 Case Study
## SailPoint IdentityIQ Engineer Exam

### YOU-CAN-SKIP-THE-NEXT-PART SUMMARY

SailPoint used SmartItems exclusively for their two certification exams published in 2018. They were attracted to SmartItems primarily for the security and cost-saving benefits. From the outset, we learned creating an exam with SmartItems was more of a design effort than a writing one. To enable this effort, item writers were divided into small teams, including a coder for each team.

During the item development process, we created a significant number of SmartItems. Code, along with other pieces of item content, were reviewed constantly and fixed when problems were discovered. Field tests of the items for the two exams went well, and a cutscore was set based on the collected data. Statistical analyses for the exams revealed the SmartItems to be high-quality items that ranged in difficulty and other properties. The large majority met the usual psychometric quality control criteria required to serve on the exams.

Since the tests went operational, the SmartItems have performed well, with the typical variability in item quality for certification exams made up of traditional items. Test-level reliability and validity evidence provided support for the use of SmartItems to produce test scores for high-stakes certification decisions.

## Introduction

SailPoint began its certification program in 2018. The mission of the program was similar to that of every other technology-based certification program: to qualify candidates that have sufficient skills in order to support the company's technology. Instead of creating exams using traditional methods—with static items and multiple forms—the SailPoint program leadership decided to utilize SmartItems. This decision was primarily made for security reasons, but also for cost-saving and other benefits described in earlier chapters. The goal for the exam development process was to create a single form entirely of SmartItems, with one SmartItem per identified skill. The skills, referred to as "objectives" by

the SailPoint certification team, were determined using standard job analysis methodology.

Originally, Sailpoint planned on creating 170 traditional MC items to support two forms of the test in a single test development workshop. Once the decision was made to use SmartItems rather than traditional items, the structure of the workshop changed. Instead of 170 items, it was instead decided to create one, perhaps two, SmartItems per objective (depending on the complexity of the objective). Since multiple forms are unnecessary with SmartItems,[1] it was also decided to create only a single test form with a target of around 65 SmartItems. If an objective needed to be weighted more heavily on the test, the SmartItem congruent with that objective could simply be presented on the form more than once.

## Methodology

### JOB ANALYSIS AND BLUEPRINTING

A job analysis was conducted by interviewing SailPoint subject matter experts. From these interviews, a list of objectives was created and then weighted for their perceived importance. There were 62 total objectives.

### ITEM DEVELOPMENT

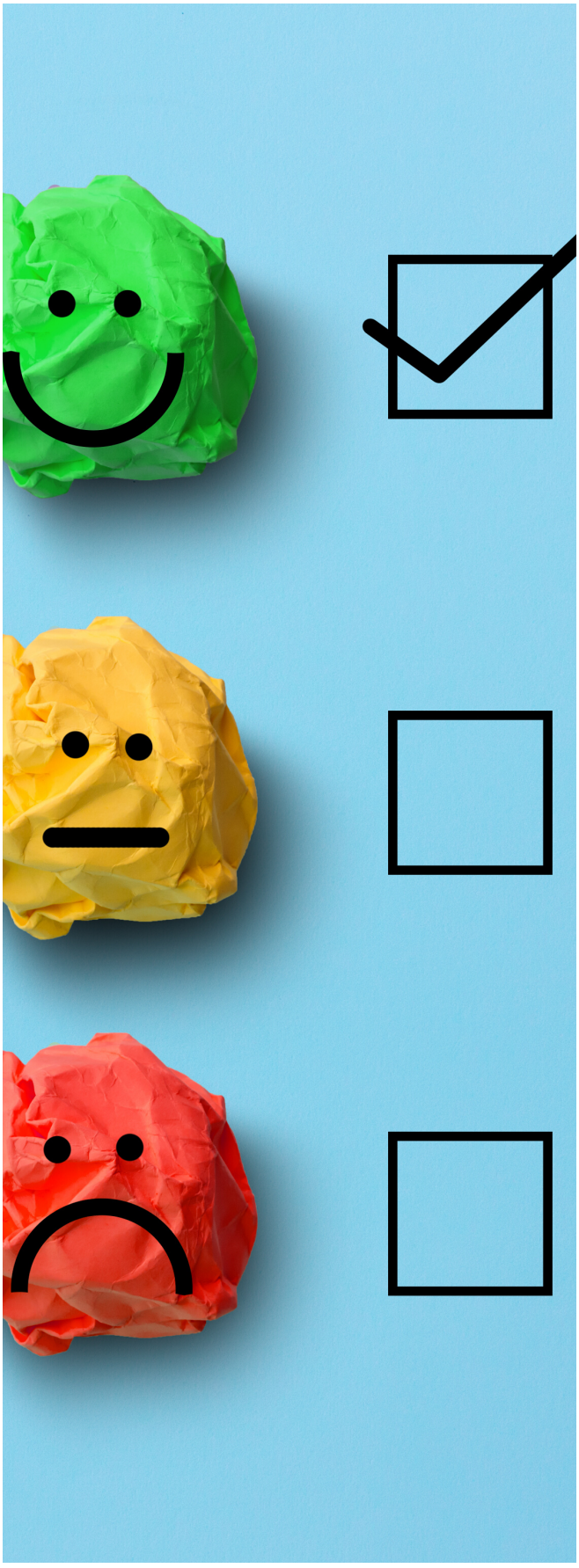
A total of 64 SmartItems were created and vetted for the objectives for the IdentityIQ Engineer exam.[2] The SmartItems used a mixture of item formats. Figure 7.0 shows the breakdown of item types used as SmartItems. To avoid the validity issues associated with testwiseness, no traditional MC items were used.

Each SmartItem was developed by a team of two or three SMEs, plus an individual with coding skills. (At the time, Caveon's GUI-based tool was still in development, so coders were necessary to develop SmartItems.) The coder provided input to the SMEs on whether the design of their item was both functional and easy to code, thereby maximizing efficiency. The coder then implemented the SMEs' design in Scorpion, Caveon's exam development and delivery platform.

| Item Types as SmartItems for Case Study #1 | |
|---|---|
| DOMC (SuperDOMC) | 27 |
| DOMC (Coded)[3] | 27 |
| Build List (Coded) | 7 |
| Matching (Coded) | 2 |
| Short Answer (Coded) | 1 |
| TOTAL: | 64 |

**Figure 7.0**
The breakdown of item types used as SmartItems for Case Study #1.

Once a draft of the SmartItem was complete, the SmartItem was "previewed"[4] a sufficient number of times to determine whether the item was functioning as designed or needed revisions. This review process differed from what is typically done for traditional items because the SmartItems appear differently each time they are rendered. This process was more similar to a software quality assurance check than to a conventional item review. SmartItems that failed to pass this quality check were discarded or revised and then previewed again. Those that passed this quality step were compiled together for a field test.



## FIELD TEST

The field test, often referred to as a "beta test" or an item "pre-test," was conducted to obtain an empirical measure of SmartItem quality. Were the SmartItems too easy or too difficult? Did they discriminate well among test takers?

Several dozen individuals were recruited to take all 64 SmartItems that made up the field test exam. The participants were recruited to span a range of ability in order to:

- Obtain stable estimates of item performance metrics, and
- Help determine the pass/fail standard for the final exam.

Based on field test results, 49 out of the 64 SmartItems were deemed to be of sufficient psychometric quality to serve on the actual certification exam. The issues with the remaining SmartItems were typical of those found in all test development projects. Some SmartItems were too difficult or too easy. A few had low correlations with the total test score.

All of the remaining 15 items were carefully reviewed and revised and then included in the certification exam as unscored questions to collect data on the new versions. It was hoped that these repaired SmartItems might eventually perform well enough to be included on the scored portion of the exam.

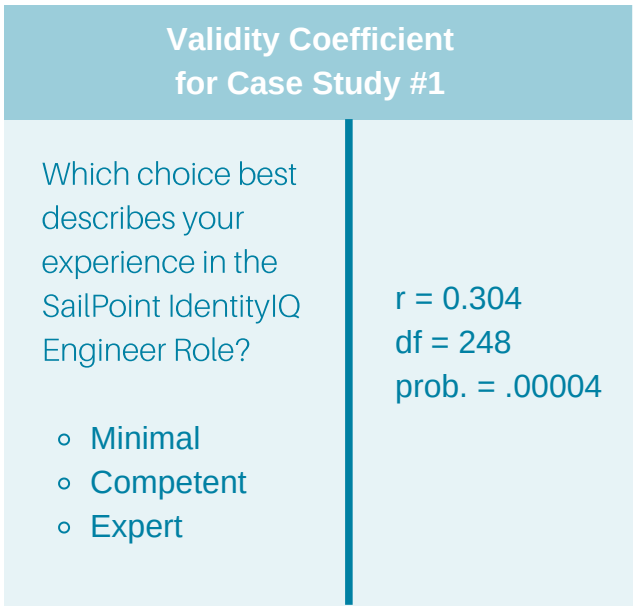



**Figure 7.1**
The survey question and its validity coefficient.

## VALIDITY EVIDENCE

The field test also included survey questions. One question asked the candidates to indicate their proficiency in the SailPoint skills covered by the exam. Their responses were then correlated with their test scores, providing empirical evidence of validity. Figure 7.1 shows the survey question and its validity coefficient.

## RELIABILITY EVIDENCE

The Cronbach's Alpha reliability coefficient, calculated using only the data from the 49 scored SmartItems, was $\alpha = .75$.

## SETTING THE CUTSCORE

The Contrasting Groups method (Zieky, 2001) was used to set the cutscore for this exam. The field test provided a way to create a cutscore. Field test participants were pre-sorted, based on judgments by the SailPoint management team and various other supervisors regarding the capability of each participant. Based on these judgments, the field test participants were divided into three groups:

1. Those who are expected to pass the exam
2. Those where it is unclear if they should pass or not, and
3. Those who should not pass the exam

After the SmartItems were evaluated and the final 49 SmartItems selected, the performance of the individuals was re-scored based on those 49 items. The score distributions for the three groups were plotted and compared. A cutscore that minimized classification errors was set.

Field test participants were told in advance that their score, based on qualified SmartItems, would be evaluated against an empirically derived cutscore following the field test. It was assumed the motivation these individuals would feel to obtain the certification would produce test-taking behavior similar to candidates who would be taking the operational exam in the future.

## Performance of the Operational Exam

### SMARTITEM DISCLOSURES

One year after the exam had been released and made available to candidates, there still had not been any public disclosures of exam content on the Internet. This was confirmed by extensive web patrolling efforts throughout the life of the exam. This result is very unusual for IT-based certification programs, as test content is usually stolen and disclosed within days.

### PSYCHOMETRIC PERFORMANCE

With over 375 tests administered as of the authoring of this case study, the SailPoint IdentityIQ Engineer exam continues to perform well. Similar to the field test results, the items show a typical large range of variability in item difficulty and item discrimination.

### VALIDITY EVIDENCE

Like the field test, the operational test also included survey questions. One of those questions asked the candidates to indicate their proficiency in the SailPoint skills covered by the exam. Their responses were then correlated with their test scores, providing initial empirical evidence of validity. Figure 7.2 shows the survey question and its validity coefficient.

| Validity Coefficient for Case Study #1 | |
|---|---|
| Which choice best describes your experience in the SailPoint IdentityIQ Engineer Role?<br>○ Minimal<br>○ Competent<br>○ Expert | r = 0.24 |

**Figure 7.2**
The survey question and its validity coefficient for Case Study #1.

### RELIABILITY EVIDENCE

The Cronbach's Alpha reliability coefficient, calculated using the data from the 55 scored SmartItems used on the operational exam, was $\alpha = .74$.

# 2 Case Study SailPoint IdentityIQ Architect Exam

### YOU-CAN-SKIP-THE-NEXT-PART SUMMARY

Case Study #2 describes the second exam published by SailPoint, this one titled the Identity IQ Architect Exam. The experience, including the positive outcomes, was similar to the first exam except for one main difference:
For this exam, we changed the SmartItem review process to improve our ability to evaluate each SmartItem. A new "mapping" phase was added to the development process. In this phase, the SMEs created the item variables and values[5] for a SmartItem and indicated the relationships between them. This helped later reviewers understand how the SmartItem functioned in order to evaluate if it is functioning properly.

## Introduction

SailPoint began the planning and development process for its second certification exam—the SailPoint IdentityIQ Architect—a bit differently than it had for the first exam. From the outset of the planning process, the goal was to create an exam completely of SmartItems. Given the experience from the initial exam, a small change was made in the SmartItems writing process.

The development team leaders further refined a step that takes place between an SME designing a SmartItem and the actual coding of the SmartItem. The refinement of this step involved the item SME designers/writers creating stronger "road maps" of the SmartItem. For each SmartItem, they would create a roadmap that included the stem, all variables, all option sets, and any extra guidance for the coder on the relationships between these components. Additionally, this "road map" was archived in the development software so anyone at any stage of the exam development process could view the original intentions of the SMEs for each item.

With one SmartItem exam written, the development team was able to provide more examples of SmartItems to help the SME designers and coders envision and create new items. This made it easier for the SMEs to learn the process and think creatively about designing items to cover a specific objective.

## Methodology

### JOB ANALYSIS AND BLUEPRINTING

A job analysis was conducted in a manner similar to Case Study #1. A total of 67 objectives were identified.

### ITEM DEVELOPMENT

Based on the objectives of the IdentityIQ Architect exam, 67 SmartItems were eventually created. The SmartItems used a mixture of item formats. Figure 7.3 shows the breakdown of item types used as SmartItems. As with the original exam, no MC items were used. Each SmartItem was developed by a team of two or three SMEs who designed and crafted the SmartItem. A coder was assigned to each team to create the items in Scorpion, Caveon's exam development and delivery platform. The writing process was enhanced with a "mapping" phase where SMEs documented their variables and logic behind each SmartItem. This additional step helped subsequent reviewers understand the intended function of the items.

**Item Types as SmartItems for Case Study #2**

| | |
|---|---|
| DOMC (SuperDOMC) | 35 |
| DOMC (Coded) | 28 |
| Build List (Coded) | 3 |
| Matching (Coded) | 1 |
| TOTAL: | 64 |

**Figure 7.3**
The breakdown of item types used as SmartItems for Case Study #2.

### FIELD TEST

As with the first case study, a field test of the 67 items was conducted. Participants were recruited whose abilities spanned a range of competency on the content comprising the IdentityIQ Architect exam. As before, the field test not only evaluated item performance empirically, but also provided the data to set the cutscore. Based on field test results, 61 of the 67 total SmartItems were deemed to be of sufficient psychometric quality to use on the actual certification exam. The remaining six SmartItems were revised and then included on the certification exam as unscored items in order to collect new data.

## VALIDITY EVIDENCE

The operational test also included survey items. Two of those survey items asked the candidates questions related to their proficiency in the content of the exam. Their responses were then correlated with their test scores, providing empirical evidence of validity. Figure 7.4 shows the survey questions and validity coefficients.

| Validity Coefficients for Case Study #2 | |
|---|---|
| Have you previously taken any IdentityIQ Implementer courses offered by SailPoint?<br>○ Yes<br>○ No | r = 0.49 |
| Which choice best describes your experience in the SailPoint IdentityIQ Architect role?<br>○ Minimal<br>○ Competent<br>○ Expert | r = 0.51 |

**Figure 7.4**
The survey questions and validity coefficients for Case Study #2.

## RELIABILITY EVIDENCE

Cronbach's Alpha reliability coefficient, calculated using only the data from the 61 scored SmartItems, was $\alpha = .82$.

## SETTING THE CUTSCORE

As with Case Study #1, the Contrasting Groups method was used to set the cutscore for this exam. Field test participants were pre-sorted based on the SailPoint management team's experience with participants. They were then divided into two groups:

1. Those who are expected to pass the exam, and
2. Those who should not pass the exam.

After the SmartItems were evaluated and the 61 final SmartItems were selected, the performance of the individuals was re-scored based on those 61 items. The score distributions for the two groups were plotted and compared. A cutscore was set that minimized classification errors.



## Methodology

What did we learn from the two case studies? What are some of the initial insights from using SmartItems on actual high-stakes information technology certification exams? Here are some initial conclusions:

1. Using SmartItems reduces initial item development costs.[6] It is also likely they will reduce long-term maintenance costs because SmartItems do not need to be replaced.
2. With the addition of a unique step or two,[7] SmartItems can be created easily in a typical item development workshop environment. Developing SmartItems takes no more time than a traditional workshop.
3. MC items, either as traditional MC items or as MC-based SmartItems, are generally unnecessary. DOMC is a preferred substitute because it performs as well or better statistically and removes the testwiseness advantages for some test takers.
4. Not all SmartItems require coding. The SuperDOMC format[8] was sufficient and appropriate for almost half of the skills on the SailPoint exams.
5. SmartItems, when created, can be effectively reviewed for accuracy, bias, etc.
6. SmartItems can use several varieties of item formats: DOMC, SuperDOMC, Build List, Matching, and Short Answer.
   a. DOMC: Some DOMC items employed code to display the item content.
   b. SuperDOMC: No code was used. Instead, a large number of options were produced.
   c. Build List: Here, examinees selected from a large list and dragged elements from that list to create another. Then examinees re-arranged the list according to the instructions in the item.
   d. Matching: This Matching format required the examinees to match two lists using drop-down menus.
   e. Short Answer: Examinees were required to type their answer in the response box provided using their keyboard.
7. SmartItems performed well statistically in both the field tests and the operational tests.
8. SmartItems, like traditional items, may be constructed poorly. Weak items are determined through field testing and statistical analysis.
9. A test completely comprised of SmartItems can produce strong evidence of reliability and validity.
10. SmartItems can be used as part of the process to set cutscores using the Contrasting Groups method.

# 07 CONCLUSION

It is clear from these two case studies that SmartItems are a viable alternative to using fixed-content items. By replacing the traditional items with SmartItems, security is enhanced and test development costs are lowered. In addition, the influence of testwiseness, benefiting some test takers to the detriment of others, is virtually eliminated when DOMC is part of the SmartItem design. The test takers adapted easily to the new DOMC format and to the variable nature of SmartItem renderings.

For SailPoint, the experience from the first two exams has encouraged the program to continue to use the innovative methodology for their third exam, published in the fall of 2019.

With SailPoint's support, SmartItem technology is no longer just an innovative idea. With these case studies, SmartItems went from a promising innovation to actual use on high-stakes tests.

## FOOTNOTES:

1. The purpose of multiple forms is to provide candidates with a different form when re-testing takes place, as it isn't advisable to provide the same form a second time. In addition, a second form would make it more difficult to cheat even if incomplete pre-knowledge were obtained from a braindump site. However, these security measures are unnecessary when SmartItems are used, as each test taker experiences a form different from everyone else's, or if he or she took the exam a second, third, fourth, etc., time.
2. When objectives describe complex skills, it may be reasonable to create more than one SmartItem for the objective. The decision to do this a couple of times resulted in the number of SmartItems (n=64) being higher than the number of objectives (n=62).
3. The term "coded" used here indicates programming code was used in the development of some SmartItems. Other SmartItems, which are always selected-response items, do not use code, but instead use a much larger number of options.
4. Previewing an item means the SmartItem was "run" in the same way it would appear and function on an exam for an actual test taker. The SmartItem reviewers would be able to answer and score the SmartItem as part of this step. By doing this a number of times with no apparent issues, the reviewers can be confident that the SmartItem would perform similarly on the operational exam.

## FOOTNOTES CONTINUED

5. An example would probably be helpful here. A variable could be "mammals" and the values would be a list of specific mammals, such as bear, dog, llama, etc. The list of values could number in the hundreds, thousands, or more.
6. Given that this case study describes one of the initial pioneering efforts, it isn't clear how much savings will be gained by additional experience in designing and producing SmartItems.
7. SmartItem development has unique steps, including a more complex design phase, along with innovative item review steps.
8. SuperDOMC items are items where the number of options, both correct and incorrect, exceeds the typical number or three to five. SuperDOMC items commonly use dozens or hundreds of options.

Chapter

# 08 EVIDENCE FROM SCIENTIFIC EXPERIMENTS

## PART 1

The last chapter described two case studies where SmartItems were used exclusively in high-stakes certification exams. The evidence from those studies indicates SmartItems performed well for the certification. In this chapter, I increase the scope and provide evidence from scientific experiments that SmartItems work—that they function psychometrically and provide the promised security benefits.

This chapter outlines the details of three scientific experiments. One experiment uses SmartItems as described in this book. The other two use a close approximation to SmartItems, supporting the increased use of randomization. These experiments will prove, to a high scientific standard, that the random variability in item content produced by SmartItems results in stable test scores and comparable item performance.

Before getting started, let's take a moment to quickly refresh our memory on the most important features of SmartItems:

1. They cover the objective completely.
2. They may or may not use code.
3. They render in more than one way (even in thousands or millions of ways).
4. Test takers experience different renderings of the SmartItem.

All the SmartItems in these experiments have these characteristics.

### Case Study vs. Experiment

What is the difference between a case study and an experiment? The difference is the extent to which the researcher controls variables and conditions. A case study does not control study conditions, while an experiment does. As a result, cause and effect cannot be determined in case studies, only in experiments. (See Chapter Seven for more details about case studies.)



An experiment uses procedures to control for extraneous influences with the express purpose of determining cause and effect. (At least within a small degree of acceptable statistical error.) Because of this, experiments are the preferred method for conducting research.

### Three Scientific Experiments on SmartItems

Three scientific experiments have been conducted on SmartItems (so far). Each Experiment is outlined below and demonstrates the value and effectiveness of randomizing test content as a tool for presenting secure, high-quality items. While all three experiments focus on the concept of content randomization, they vary in the amount of random content that is administered to test takers.

The first experiment directly compares SmartItems. These SmartItems are presented in two formats, and can be directly compared to traditional multiple-choice items. The second experiment compares MC with DOMC, focusing on the randomization of DOMC options. The third also compares MC to DOMC, but increases the number of options available in each DOMC item option pool, thereby increasing the random variability of content. These expanded DOMC items are called "SuperDOMC." They can be true SmartItems when the available options enable the item to cover the entire breadth of the objective. (Remember, this is one of the key characteristics of what makes a SmartItem. See Chapter Two for more details.)

The results from all three experiments clearly show that items utilizing randomization, when constrained by the skill description, had the same end-of-test scoring outcomes as traditional MC items. They also show that randomization allows for better control of item exposure, greatly reducing (and even eliminating) forms of cheating and test theft.

## The Experiments:

The information provided for each experiment is divided into four sections: an introduction, a description of the research design, the results, and a conclusion. This information will show how empirical research supports claims on the quality and effectiveness of SmartItems.

# Scientific Experiment 1
## Comparing the Performance of DOMC SmartItems, MC SmartItems, and Traditional MC Items

### YOU-CAN-SKIP-THE-NEXT-PART SUMMARY

In this research experiment, the performance of DOMC-based SmartItems was compared directly with (1) MC-based SmartItems and (2) traditional static MC items. The content for all three item types was developed from the same objectives; the significant difference was the item format. Looking at individual items, MC items were overall easier than DOMC items. This effect was expected and is likely due to the use of testwiseness in answering MC questions compared to DOMC questions. In addition, there were a few idiosyncratic differences in some items where item content interacted with item format.

Most importantly, results from the experiment showed item format had no effect on reliability and validity measures. This finding supports the use of SmartItems on high-stakes tests.

## Introduction

Here are a few of the critical questions about SmartItems that were addressed in this experiment:

1. Performance: Do SmartItems perform well on an exam according to psychometric standards, despite each question's random renderings? Will the SmartItems discriminate between individuals who are qualified and those who are not? Can statistics such as the point-biserial correlation statistic (from classical test theory) or the *a* parameter (from IRT) confirm SmartItems' performance?
2. Difficulty: Are SmartItems appropriately difficult? Items that are too easy or too difficult contribute relatively little information useful to a test score. Can psychometric measurements be evaluated using the p-value (proportion of individuals answering the SmartItem correctly) or the *b* parameter (if using IRT statistics), encourage the use of SmartItems on a typical fixed-length exam?
3. Comparison: Given the newness of SmartItems, how might their performance with traditional items compare in regard to item statistics?
4. Format: How will SmartItems using the DOMC format compare with SmartItems utilizing the MC format?



To answer these questions, it was important to choose a content area and a set of participants where the likelihood of cheating was expected to be low. For that reason, the test used in this experiment can be considered to be low-stakes or no-stakes. The content area was Season 1 of *Game of Thrones* (GoT), an HBO television series. This content was chosen because it allowed the use of a number of large fan forums, or large communities of individuals interested in the same topic. The item content consisted of two types of questions in equal number:

1. Recall questions: Items that asked the research subjects to recall facts about GOT Season 1.
2. Reasoning questions: Questions that asked the research subjects to use reason or logic to determine the answer.

The test scores were not used for any high-stakes decisions, and personal identifying information was not requested or collected. The "fans" were asked to take the test without cheating, and a large majority (about 96%) reported that they answered honestly. Any participant who reported using outside assistance to answer questions had their data removed from the analyses.

## Method

### Subjects

Participants were volunteers recruited from several online GoT fan forums. The subjects were recruited by providing a request for help on a research project and a link to the exam. Here is the initial post:

> Consider yourself a true *Game of Thrones* fan? Never met a quiz you couldn't beat? Well, you have finally met your match!
>
> https://sei.caveon.com/launchpad/game-of-thrones-season-1?fbclid=IwAR2vSSSga5b4TaUuiCAfTEAMi_NEhp3apasRlV08nJgpsmVvnSdNstbkYnc
>
> I've created this devilishly tricky quiz on the SEASON 1 OF *GAME OF THRONES* as part of a research project. Click the link to take the test, see how you do, and help me with my research at the same time!
>
> Thanks for your help!

The link in this announcement remains active. You are welcome to go to it and experience the same exam as this experiment's participants.

Some participants took the test multiple times. All data except for a participant's first time were excluded from the analyses. Additionally, data were excluded from participants who:

- Answered less than half of the items,
- Took much longer than the expected time on the test (e.g., more than 30 minutes for a test where most participants took about 10 minutes), or
- Took the test much faster than the other participants (less than three standard deviations of the mean time).

The total number of participants was 1,156. Several participants were excluded based on the criteria described above, leaving the data of 1,031 participants for analysis.

## Item Design

The 21 SmartItems were created by a GoT subject matter expert and were each congruent with 21 different objectives or skills.

> Sample Objective: Recall the correct order of appearance of characters in Season 1 of *Game of Thrones*.

There are 35 characters in Episode 1 of *Game of Thrones*, and the multiple-choice SmartItem for this skill can render in 1,681,680 ways. The SmartItem DOMC for this skill can render in 2,704,702 ways.[1] For both item formats, different option orders were considered different renderings.

Sample multiple-choice SmartItem rendering:

> Which options shows the correct order of appearance between two characters as they appear in Episode 1 of Season 1?
>
> - Joffrey Baratheon before Myrcella Baratheon
> - Myrcella Baratheon before Gared
> - Robert Baratheon before Joffrey Baratheon
> - Joffrey Baratheon before Gared
> - Viserys Targaryen before Myrcella Baratheon

## Variability of SmartItems and Traditional Items

Figure 8.0 shows the variability of all 63 items in the experiment. Options were presented in a random order. Even the fixed traditional MC items, while static in content, did appear in 120 different ways to participants because the options varied in order. The variability of the SmartItems depended on the number of coded variables and the number and order of options presented. Figure 8.1 shows the variability for the SmartItem DOMC items, the SmartItem MC items, and the traditional MC items.

| Item | MC | SmartItem MC | SmartItem DOMC |
|---|---|---|---|
| 1C | 120 | 158400 | 254760 |
| 10C | 120 | 480 | 772 |
| 21C | 120 | 1200 | 1930 |
| 2C | 120 | 600 | 965 |
| 11C | 120 | 277200 | 445830 |
| 3C | 120 | 3720960 | 5984544 |
| 4C | 120 | 18720 | 30108 |
| 17C | 120 | 25200 | 40530 |
| 12C | 120 | 1681680 | 2704702 |
| 5C | 120 | 16800 | 27020 |
| 13C | 120 | 360 | 579 |
| 19C | 120 | 840 | 1351 |
| 16C | 120 | 7200 | 11580 |
| 14C | 120 | 600 | 965 |
| 15C | 120 | 480 | 772 |
| 6C | 120 | 4080 | 6562 |
| 7C | 120 | 360 | 579 |
| 8C | 120 | 50400 | 81060 |
| 9C | 120 | 153600 | 247040 |
| 18C | 120 | 480 | 772 |
| 20C | 120 | 480 | 772 |

**Figure 8.0**
The variability of all 63 items in the experiment.

## Test Design

The experimental design counterbalanced the three item formats across three test forms[2]: Test Form A, Test Form B, and Test Form C. Figure 8.1 shows these conditions.

The SmartItems, both MC and DOMC types, were rendered on-the-fly according to their design and programming. The MC items were fixed and did not change during the study, except for the random ordering of the options. For each of the 21 objectives, items in three formats were created by the GoT SME:

- One SmartItem with DOMC response format (a total of 21 SmartItems)
- The same SmartItem with MC response format (a total of 21 SmartItems)
- A fixed MC item. This item was selected by rendering the SmartItem MC once and using that rendering as the fixed MC item. It remained in this fixed form for the duration of the study. (A total of 21 fixed MC items.)

| 21 ITEMS | TEST FORM A | TEST FORM B | TEST FORM C |
|---|---|---|---|
| 7 Items | SmartItem DOMC | MC | SmartItem MC |
| 7 Items | SmartItem MC | SmartItem DOMC | MC |
| 7 Items | MC | SmartItem MC | SmartItem DOMC |

**Figure 8.1**
The experimental design counterbalances the three item formats across three test forms: Test Form A, Test Form B, and Test Form C.

The content of the items was consistent across the three formats. The SmartItem DOMC and the SmartItem MC differed in three ways:

1. How the selected-response options were presented (one at a time versus simultaneously)
2. How the questions were answered (selecting if each option is correct or incorrect versus selecting a single option from a list of options)
3. How the items were scored (correctly identifying each presented option as correct or incorrect versus correctly selecting an option from a list).

If the difference between MC and DOMC is unclear, please refer to Foster & Miller (2009) or visit the site www.trydomc.com.

Randomization was used in four ways in this experiment:

1. One of the three forms was randomly assigned to each participant.
2. The order of the 21 items within each form was randomized for each participant.
3. The order of the options for all three formats (SmartItem DOMC, SmartItem MC, and MC) was randomized for each participant.
4. The 21 items of each format were randomly assigned to forms, with each form consisting of exactly seven items of each format.

The counterbalanced design and the use of randomization to control for known and unknown extraneous variables permitted the direct, unconfounded comparison of item formats.

The dependent measures used in the study were the responses made by the participants and the latencies for those responses.



## Survey

A nine-question survey was given to participants at the end of the exam. Responses to the survey questions were analyzed and correlated with test scores.These are the survey questions given to participants:

1. What is your age?
    - Participants typed in the numbers representing their age.
2. To which gender do you identify?
    - Participants chose between "Female," "Male," "Other," and "Prefer not to answer."
3. To which racial or ethnic group do you most identify?
    - Participants chose between "Asian," "Black or African American," "Caucasian or White," "Hispanic or Latino," "Native American or American Indian," "Pacific Islander," "Multiracial," "Other," or "Prefer not to answer."
4. Please rate your ability at taking multiple-choice tests.
    - Participants responded by choosing "Excellent," "Very Good," "Good," "Poor," or "Very Poor."
5. Which testing method would you prefer?
    - Participants chose between "A testing method that removes the ability to use test-taking skills (e.g., comparing options, eliminating unlikely answers, etc.) so everyone can be judged equally according to their knowledge of the question's content," or "A testing method that gives good test takers, or people with lots of experience taking tests, an advantage."
6. With which statement do you most agree?
    - Participants chose between "I prefer to answer DOMC questions," "I prefer to answer traditional multiple-choice questions," or "I have no preference."
7. How would you rate your knowledge of *Game of Thrones*, Season 1?
    - Participants chose between "Beginner," "Intermediate," and "Expert."
8. How many times have you attempted this test?
    - Participants chose between "This is my first attempt," "2 times," "3-5 times," "6-10 times," or "More than 10 times."
9. Did you use any external help to answer these questions (e.g., looking up answers on the Internet, asking a friend for help, etc.)?
    - Participants chose between "Yes" and "No."



## Procedures

Participants for the experiment were recruited through a public post sent out to a fan forum for GoT. The post included a link to take the exam (See earlier in this chapter). The participants were required to complete a typical reCaptcha security task, and were asked to read information about the exam and to follow instructions. Participants were not made aware of the SmartItem conditions, but were told they would see both MC and DOMC items.

Additional instructions disclosed the number of questions they would see and be asked to answer (21). They were told "It should only take about 10 minutes to complete the quiz." They were also asked not to seek help from other resources or other persons. Both MC and DOMC items were described in detail, along with what each format entailed. This was followed by a request to respond to a "brief survey." Finally, they were reminded that their performance was anonymous.

As the participants answered an exam or survey question, their response and response time (amount of time it took to answer the question) were immediately recorded. When the experiment concluded, 1,156 subjects had successfully completed both the exam and the survey. Only the results from first-time test takers were included. In addition, if the participant responded to Survey Question #9 that they had received outside help, the data from that test were discarded. Of the 1,156 participants, data from 1,031 remained after exclusions (More details above). The remaining data were then subjected to statistical analysis.

The analyses included:

- Comparison of Forms
- Calculation of a Reliability Coefficient
- Calculation of a Validity Coefficient between total score and Question #7
- Calculation of Item Performance Metrics

## Analyses

Comparison of Forms—The three forms were designed to perform equivalently. Each form consisted of 21 items and was equally balanced in terms of item format and content. Nevertheless, since 14 of the 21 items on each form were SmartItems, there was some uncertainty as to whether the random variability of those items would impact overall form performance. Figure 8.2 shows the score distribution for each form.

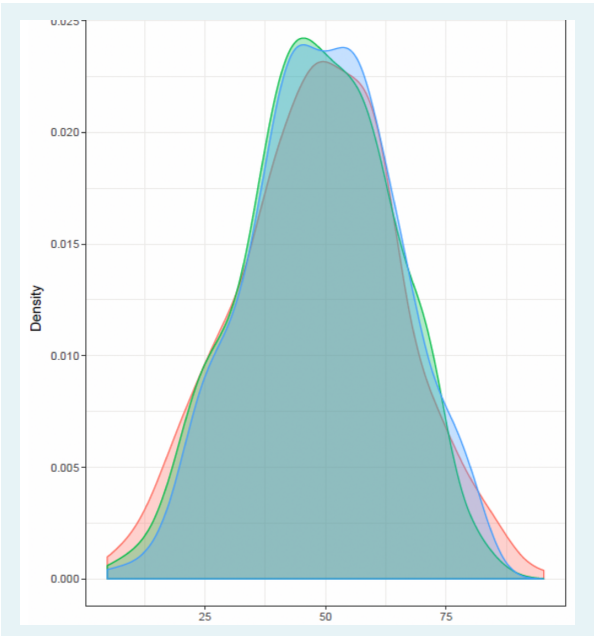



Figure 8.2
Score distribution by form.

The mean scores for the three forms were Form A = 47.79, Form B = 48.94, and Form C= 47.62. Additional comparisons of the forms indicated no differences in participant demographics (age, gender, and race/ethnicity), in reliability, or in validity (correlation coefficient between Survey Question #7 and test score).

Forms A, B, and C can be considered to be equivalent.

Reliability—The question could be asked, "With all of the variability associated with the random renderings of SmartItems, should the reliability be degraded compared to static items?" Reliability coefficients were calculated for all three formats. These can be seen in Figure 8.3.

Figure 8.3

Reliability coefficients (Cronbach's alpha) and SEM for each item format by test form.

| 7-Item Subtests by Form | N | Predicted Alpha with 70 Items | 7-Item Alpha | SEM |
|---|---|---|---|---|
| SmartItemDOMC (Form A) | 360 | .48 | .08 | 1.21 |
| SmartItemDOMC (Form B) | 319 | .81 | .29 | 1.22 |
| SmartItemDOMC (Form C) | 354 | .85 | .36 | 1.17 |
| | | | | |
| SmartItemMC (Form A) | 360 | .85 | .37 | 1.19 |
| SmartItemMC (Form B) | 319 | .88 | .42 | 1.16 |
| SmartItemMC (Form C) | 354 | .74 | .23 | 1.20 |
| | | | | |
| Fixed MC (Form A) | 360 | .80 | .28 | 1.11 |
| Fixed MC (Form B) | 319 | .63 | .14 | 1.20 |
| Fixed MC (Form C) | 319 | .88 | .43 | 1.15 |

A review of the values in Figure 8.3 reveals there are no consistent differences in reliability due to item format. However, there is some variability in the reliability coefficients. This is likely due to using the results of only seven items to calculate each coefficient. If a test were 70 items in length instead of seven, the predicted alpha value indicates that for all three formats, lengthier tests would have stronger reliability. The reliability of the test is not affected by the random variability of SmartItem content

Validity—Validity is a multi-dimensional concept extending beyond what might be presented in this section. For example, SmartItems cover the relevant skill completely. As such, from the perspective of SMEs, SmartItems are logically valid—even perfectly valid—relative to the skills measured by the test.

However, another way to gather validity evidence from a research study is to compare the scores on a test with another criterion. In this case, the criterion variable is the responses participants made to Survey Question #7. It states, "How would you rate your knowledge of *Game of Thrones*, Season 1?". Participants chose between "Beginner," "Intermediate," and "Expert." It is logical to expect those who answered "Expert" would obtain higher mean scores than those who answered either "Beginner" or "Intermediate." In addition, a Pearson correlation statistic comparing the responses to Question #7 to test scores should be positive.

Figure 8.4 shows the mean scores for the survey responses given. Figure 8.5 shows the correlations comparing survey responses to total test score (based on 21 items) to item format scores (based on seven items). All correlations were significant ($p < .001$).

| GOT Knowledge | *N* | Mean Score | SD of Score |
|---|---|---|---|
| Beginner | 160 | 8.06 | 2.96 |
| Intermediate | 652 | 9.99 | 3.05 |
| Expert | 219 | 11.91 | 2.50 |

Figure 8.4
Means scores and standard deviations of test form scores.

For test scores (21 items)
$r(1,029) = 0.37, p < .001$

For MC scores (7 items)
$r(1,029) = 0.30, p < .001$

For SmartItemDOMC (7 items)
$r(1,029) = 0.21, p < .001$

For SmartItemMC (7 items)
$r(1,029) = 0.29, p < .001$

Figure 8.5
Correlations of responses to Survey Question #7 and both test form score (21 items) and item format score (seven Items).

Even though both DOMC-based SmartItems and MC-based SmartItems render randomly in thousands or even millions of ways, the validity evidence is strong and statistically significant. It is not different from what is obtained from fixed MC items.

Item Performance—Most test development efforts result, eventually, in data being collected from the administered items. Those data can be analyzed statistically to evaluate each static item. In this case, the data can be used to statistically analyze each SmartItem.

All 63 separate items were analyzed using classical test theory methods, producing estimates of difficulty, discrimination, and response time. Difficulty is represented by the p-value. Discrimination is represented by the corrected item total correlation (CITC). The standard deviations for those means are given within the parentheses. The means and standard deviations for these metrics are displayed in Figure 8.6.

| Format | Mean P-Value (SD) | Mean CITC (SD) | Mean RT (SD) |
|---|---|---|---|
| Fixed MC | 0.48 (.22) | 0.19 (.15) | 26.10 (11.21) |
| SmartItem DOMC | 0.43 (.18) | 0.16 (.13) | 25.20 (9.34) |
| SmartItemMC | 0.53 (.19) | 0.21 (.11) | 26.25 (10.36) |

Figure 8.6. Mean items statistics summarized by item format.

Overall, the average performance of items is not different across the three formats. However, when looking at the p-values, CITCs, and response times of specific items, certain similarities and differences are apparent. The next three figures, Figures 8.7, 8.8, and 8.9, show these comparisons visually. For each figure, the items are ranked by item difficulty based on the Fixed MC item performance. This decision was arbitrary as the difficulty values for either of the other two formats could have easily served.

Figure 8.7 compares the difficulty (p-value) for each item across each of the formats. Is one format more difficult than another for a particular item? Given how items are created, whether or not they are SmartItems, it is expected that idiosyncratic differences might interact with item format. To better read this graph, begin at the far left (the most difficult item) and then scan vertically to compare all three formats. For example, item #8 is the most difficult, at least in MC format. The SmartItem MC version is the easiest, with the SmartItem DOMC in between them. This difference does not hold across all item formats and may even reverse (see items #19).

By visually inspecting each item and the chart as a whole, certain patterns emerge. For example, the two SmartItem formats align more to each other than to the Fixed MC. In general,

especially at the moderate and easy difficulty levels, the SmartItem DOMC format appears to be the most difficult.

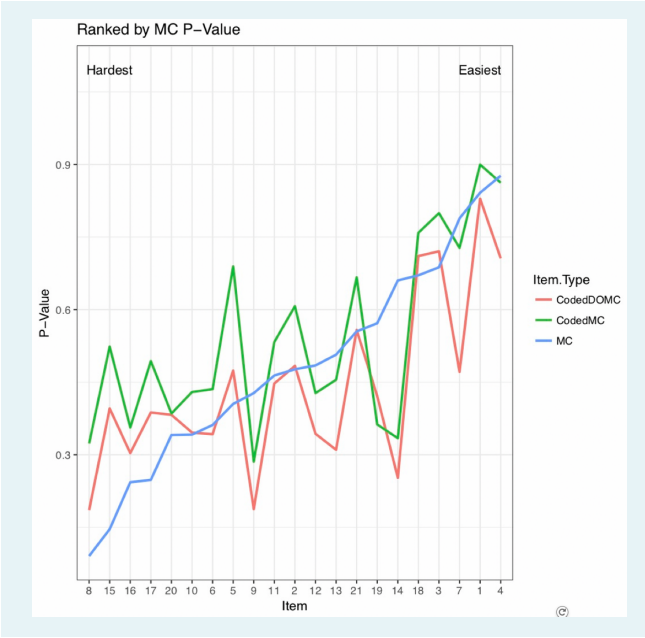



Figure 8.7. Comparison of item difficulty for each item (shown ranked by difficulty).

The patterns seen in Figure 8.7 are reflected statistically in the correlations between them. The correlation in item difficulty between MC and SmartItem MC is fairly high (r = .68). The correlation between Fixed MC and SmartItem DOMC is similar (r = .70). The correlation between SmartItem MC and SmartItem DOMC is the highest. (r=.94).

Figure 8.8 compares the discrimination values for each item across each of the three formats. Is one format more discriminating than another for a particular item? Given how items are created, whether or not they are SmartItems, it is expected that idiosyncratic differences might interact with item format. To better read this graph, begin at the far left (the most difficult item) and then scan vertically to compare all three formats. For example, item #8 is the most difficult, at least in Fixed MC format. From the graph, we see the Fixed MC version is the least discriminating, with the CITC actually being negative. The other two SmartItem formats are higher, but are still quite low. In general, there is a trend upward in the CITC as items get easier. However, items #9, #21, and others buck that trend.

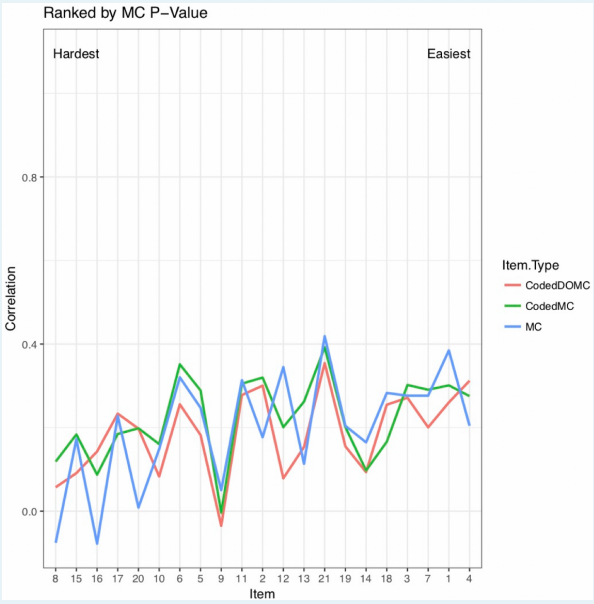



Figure 8.8. Comparison of CITC for each item (shown ranked by difficulty).

The patterns seen in Figure 8.8 are only slightly different than those for Figure 8.7, as reflected statistically in correlations between them. Despite the fact that SmartItems render variably in thousands or millions of ways, correlations between the formats are fairly high. The correlation between Fixed MC and SmartItem MC is r = .69. The correlation between Fixed MC and SmartItem DOMC is similar (r = .60). The correlation between the two SmartItem formats is r = .79.

Figure 8.9 compares the duration of the item (as measured by average response time or latency) for each item across each of the formats. Does one format require a response that takes longer than another for a particular item? Given how items are created, whether or not they are

SmartItems, it is expected that idiosyncratic differences in response time might interact with item format. To better read this graph, begin at the far left (the most difficult item) and then scan vertically to compare all three formats.

For items #15, the format that takes the most time is the MC format, followed closely by the SmartItem MC. The SmartItem DOMC takes less time. That difference is starkly different for items #20. For those items, the SmartItem DOMC takes a lot longer to answer, while the other two formats are essentially equivalent. Scanning the graph, other obvious differences between the formats are clear. While more in-depth research will help to understand this effect, it is likely the timing similarities and differences depend to a great degree on item content rather than item format.

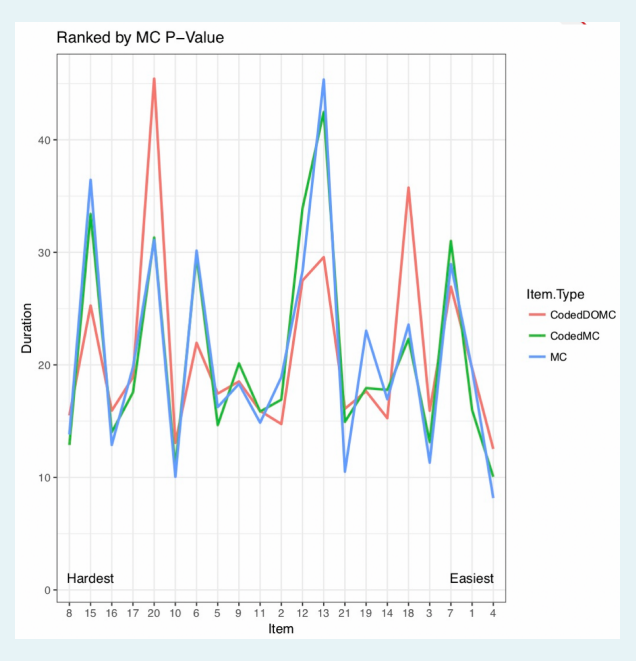



Figure 8.9. Comparison of item duration for each item (shown ranked by difficulty).

In other words, the time it takes to read the stem and compare the options, and to make a response, may depend heavily on how long it takes to read and compare the options.

The patterns seen in Figure 8.9 are very different than in the previous graphs, as reflected statistically in correlations between them. Correlations between them remain fairly high, but it is clear visually that the strongest correlation in response time is between Fixed MC and SmartItem MC (r = .93), this despite the great difference in item design between them. The correlation between MC and SmartItem DOMC is high, but not as high (r = .72). The correlation between the two SmartItem formats is similar (r = .77).



# 08 CONCLUSION

The various analyses for the experiment's results indicate that all three formats perform well and almost seem interchangeable. At the very least, there seems to be no reason to accept or reject one over the other simply on the basis of these analyses. The idiosyncrasies noted in Figures 8.7, 8.8, and 8.9, when looking item-by-item, would argue that item format may interact with specific content in such a way as to convince us further improvements can be made, and difficulty, discrimination, and response time can be better understood and managed.

The overall finding from this experiment can be described best as this: A large amount of randomization, appropriately applied as part of the SmartItem approach, can contribute to strong psychometric performance on any type of exam. SmartItems can and should (based on their ability to shut down test theft) be considered for pilot and research projects as the first step used in securing operational exams.

Although the spectrum of potential test-security risks ranges from low to high, SmartItems can viably ameliorate threats at all levels. Indeed, our findings in this experiment demonstrate their potential in mitigating a wide array of potential hazards. It can further be concluded that the decision to consider the use of SmartItems has no downside. They address major security risks without sacrificing psychometric standards or quality.

## FOOTNOTES

1. MC questions display all of the five options at one time. That can vary a bit by presenting those items in different orders. On the other hand, DOMC items don't often present all of the five options for items. If the item is answered correctly or incorrectly, the item can end without all of the options being presented. This is the reason a SmartItem DOMC with five options can vary more than a SmartItem MC item.
2. The items were randomly assigned to forms at the item level, not the block level. For example, the seven MC items on form A are NOT one particular format on Form B; they are randomly split across the formats for Forms B and C.

Chapter

# 09 EVIDENCE FROM SCIENTIFIC EXPERIMENTS

## PART 2

As was discussed in Chapter Eight, one of the most pressing questions about SmartItems is whether they perform according to psychometric standards. Do SmartItems perform well psychometrically, despite rendering randomly in so many ways? The scientific experiment described in that chapter attempted to address these questions, and it had encouraging results.

This chapter will build upon the conclusions from Chapter Eight. We will look at the results of two more scientific experiments that probe the question of whether the random nature of SmartItems impacts psychometric performance. The two research experiments presented here compare the use of traditional multiple-choice, DOMC questions, and SuperDOMC questions.

You are by now familiar with DOMC, but let's take a quick moment to discuss its variation—SuperDOMC—and its relation to SmartItems.

The SuperDOMC item is one step more complicated than the typical DOMC item. The simplest version of DOMC is a conversion from the content of an MC item, using the same stem and number of options. With a SuperDOMC item, the options are extended. Rather than the typical three to five options found in typical MC items, SuperDOMC items can have dozens—or even hundreds or thousands—of options.



As explained in an earlier chapter in this book, SmartItems do not always require code to enable them to cover an objective completely. SuperDOMC items, with a sufficient number of options and the proper content, will be able to cover the breadth of an objective. Meeting these criteria, a SuperDOMC item could also be a SmartItem.

SuperDOMC items can render in thousands of ways, simply by expanding the option pool. An example will be helpful. Consider this skill description:

> "The student can identify mammals and non-mammals."

There are thousands of potential correct options for this question, and even more potential incorrect options. So long as the skill description is worded without constraints, then common (e.g., bear and dog) and rare (e.g., platypus and ibex) mammals can be included in the list of correct options. This lack of restriction would apply to non-mammals as well. There are more than 5,000 mammal species. The number of non-mammal species is not known, but estimates given are above 2,000,000. By simple addition, the number of total options in the option pool is greater than 2,005,000.

Calculating the number of combinations of options that can be presented in a DOMC item that shows a maximum of five options, we get a value $2.18^{30}$ renderings! If the skill description can be qualified in a way similar to this, "The student can identify common mammals and non-mammals," then the maximum number can be decreased significantly, but it will still be very large.

Remember, such a SuperDOMC item is also a SmartItem because it is congruent with the skill description.



# Scientific Experiment 2
## SuperDOMC Items Compared to MC Items

**AUTHOR'S NOTE:**

This description does not present all of the details and results of the study. It does provide enough information for the reader to judge the scientific quality of the study, as well as interpret the results pertinent to understanding how SmartItems work.

To understand the psychometric properties of SuperDOMC items, it was important to choose a content area and a set of participants where cheating was unlikely. For that reason, the test used in this experiment had no stakes for the participants. There were no decisions based on the test, and personal identifying information was not collected. Similar to Experiment #1, the test for this experiment was given to individuals from online fan forums. The "fans" were asked to take the test without cheating, a request that a large majority adhered to. The few who used some sort of assistance or outside resource indicated so in a survey question. Their data were removed.

### Introduction

The content of the exam was trivia from the *Harry Potter* books by J. K. Rowling. This content was chosen because it gave us access to a number of large fan forums whose members would readily respond to such a test. The content of the items consisted of (1) questions having the research subjects recall facts from all seven books, and (2) a similar number of questions measuring reasoning skills.

This experiment compared the psychometric performance of SuperDOMC items with their MC counterpart. The stem for each question remains the same, with the only difference being the number and content of incorrect options in the option pool for each SuperDOMC item. If the random error introduced by the SuperDOMC items in the form of different sets of options seen by participants, then the only differences found should be those affected by the additional options.

Details about the design of the experiment, the participants, and the test are covered below in the Method section.

## Method

### Subjects

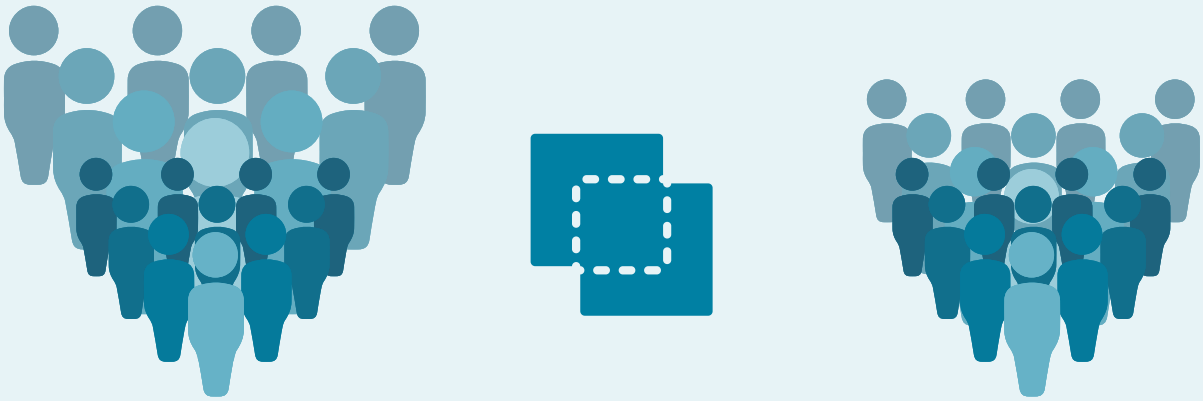


Participants were volunteers from several *Harry Potter* fan forums, and were recruited through an online post that included a link and a request for help on a research project. With a process similar to that used to recruit participants for the *Game of Thrones* forums, the *Harry Potter* forum members were offered the chance to assess their knowledge of the *Harry Potter* books. If they voluntarily chose to participate, they simply clicked on this link:

https://sei.caveon.com/launchpad/harry-potter

As with the *Game of Thrones* link, clicking on it began the process of taking the *Harry Potter* test. The link above is still active; the reader is welcome to visit the link and experience the brief exam, just as the participants did.

1,014 individuals participated in the study. 130 of those had their data excluded for various reasons (i.e., admitting to getting help with answering the questions or taking the test more than once). 884 individuals remained—first-time test takers who reported that they did not receive help answering the questions. The data from those 884 participants were used in the analyses described below.



## Item Design

The 21 SuperDOMC items were created by a *Harry Potter* subject matter expert. They were not created to correspond to particular objectives, so there is no way to classify them strictly as SmartItems. That said, they share important points in common, such as rendering in thousands or millions of ways.

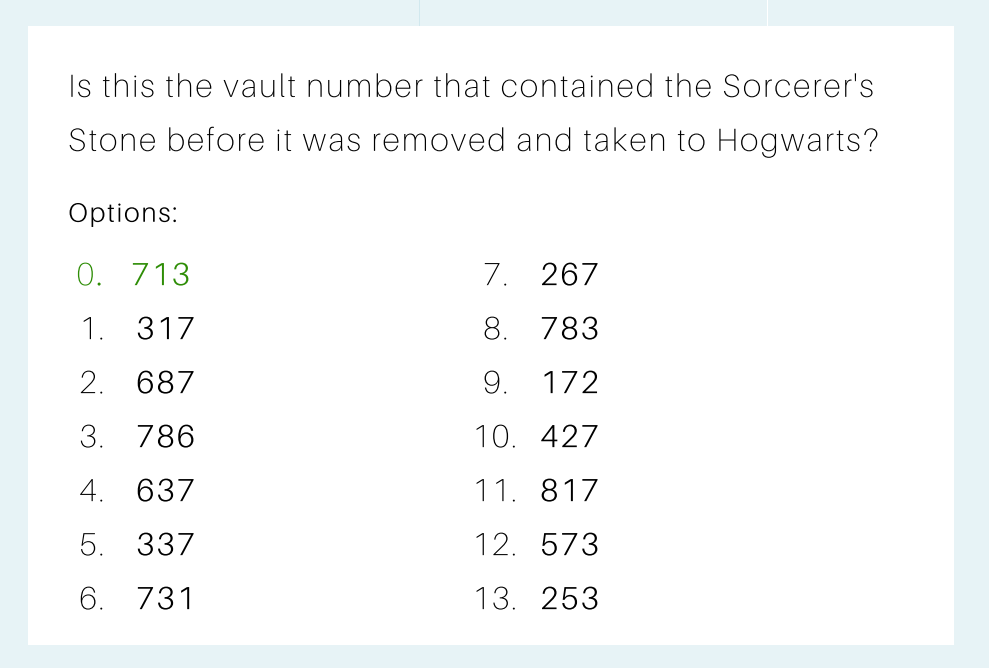



Figure 9.0
A sample of an actual SuperDOMC item used in this experiment. The reader should remember test takers will not see all of the options for this item during the test.

The SuperDOMC item pictured in Figure 9.0 can render in 1,873,872 ways, considering the option pool consists of one correct answer and 13 incorrect ones. In this study, the number of correct options was always one, and the number of incorrect ones ranged from 4 to 13. The items had a mean number of options of 8.33.

The MC items consisted of the same correct answer for each SuperDOMC item, as well as a fixed number (four) of the incorrect options.

## Test Design

The experimental design counterbalanced the two item formats across two forms: Form A and Form B. Figure 9.1 shows these conditions.

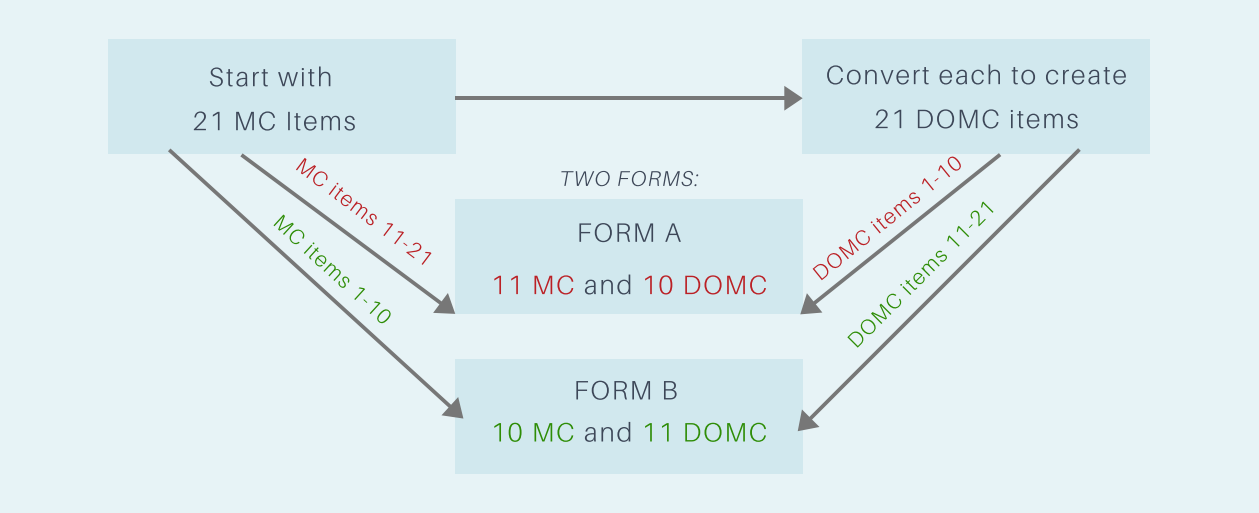



Figure 9.1
The creation of Form A and Form B. If the difference between MC and DOMC remains unclear, please refer to Miller & Foster (2009) or to the website www.tryDOMC.com.

The DOMC items referred to in Figure 9.1 are SuperDOMC items. Randomization was used in four ways in this experiment:

1. The forms were randomly assigned to each participant who volunteered to complete the exam.
2. The 21 items within each form were randomly presented to each participant.
3. The options for both varieties of items (SuperDOMC and MC) were randomly presented.
4. The 21 items of each format were randomly assigned to forms.

The counterbalanced design and the use of randomization to control for known and unknown extraneous variables permits the direct and unconfounded comparison of formats.

The dependent measures used in the study were the responses made by the participants and the latencies for those responses.

## Survey Questions

An eight-question survey was given at the end of the exam. The first three questions asked the participants for their age, gender, and race/ethnicity. The next three asked about the format of the questions, MC or DOMC. The seventh question asked them to rate their expertise relative to *Harry Potter* knowledge. The eighth survey question asked how many times they had taken this test.

Answers to the seventh question were used in the analyses described below to provide validity evidence.

## Procedures

Participants were recruited by a post to several *Harry Potter* fan forums. The post included a link to take the exam (The same link as provided above, https://sei.caveon.com/launchpad/harry-potter). The volunteers were required to complete a typical reCaptcha security task, and were asked to read about the exam and to follow instructions. Participants were not made aware of the SuperDOMC conditions, but were told they would see both MC and DOMC items.

Further instructions disclosed the number of questions (21) they would see and be asked to answer, and revealed they would have 30 minutes to complete the exam. They were informed they would not be able to return to previously viewed and answered items. They were also asked not to seek help from other resources or other persons.

Both MC and DOMC items were described in detail, along with what to expect with each question type. This was followed by a request to respond to a "brief survey" at the end of the test.

Finally, they were reminded their performance was anonymous.

While the participants answered each exam or survey question, each of their responses was immediately recorded, along with the time it took to select the response (also known as "response time" or "response latency"). When the experiment concluded, more than 1,050 subjects had successfully completed both the exam and the survey. These results are from first-time test takers only. If the participant indicated by responding to Survey Question #9 that they had used outside help, those data were also discarded. The results were then subjected to statistical analysis, which is presented in the following section.

The analyses included:

- Reliability
- Validity
- Item Performance

## Reliability

With all of the variability associated with the random renderings of the SuperDOMC items, would the reliability of the test be degraded compared to a test of entirely static items? Cronbach's alpha reliability coefficient for the MC items was 0.78. For the SuperDOMC item, the coefficient value was .76. Despite the variability difference, the reliability was unaffected.

## Validity

One source of validity evidence is to compare the scores on a test with an independent criterion. In this case, that criterion variable is an individual's response to Survey Question #7. It states, "Please rate your *Harry Potter* expertise." Participants chose between "Beginner," "Below Average," "Average," "Above Average," and "Expert." With regard to validity, it is logical to expect those who answered "Expert" would obtain higher mean scores than those who answered that they possessed lower levels of expertise. In addition, a Pearson correlation statistic comparing the responses to

test scores should be positive. The mean test scores for those five levels of expertise are shown in Figure 9.2 for both the MC and the SuperDOMC items.

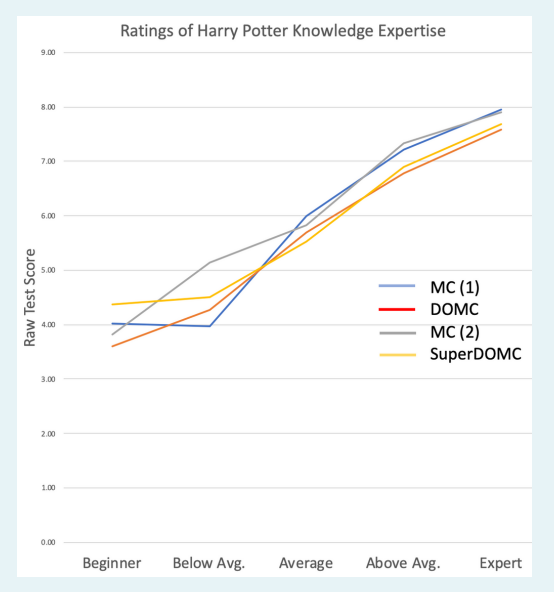



Figure 9.2. Mean test scores for both MC and SuperDOMC questions

The correlations of response to the survey question and test score for MC was 0.42, and 0.40 for SuperDOMC items. Even though SuperDOMC items render randomly in thousands or even millions of ways, especially compared to static MC items, validity evidence is strong and equivalent for both SuperDOMC and static MC items.

## Item Performance

Most test development efforts eventually result in data being collected from the items administered to test takers. Those data can be analyzed statistically to evaluate each item.

All 42 separate items were analyzed using classical test theory methods, producing estimates of difficulty and discrimination, as well as response time. Difficulty is represented by the p-value and discrimination by the corrected item total correlation, or CITC.

| Format | P-Value | CITC | Latency |
|---|---|---|---|
| MC | .67 | .34 | 11.09 |
| SuperDOMC | .63 | .33 | 14.49 |

Figure 9.3
Mean Item Statistics Summarized by Item Format

DOMC items, including SuperDOMC items, are usually perceived as more difficult, and that difference is usually a few percentage points as indicated by a lower p-value. DOMC-based items remove most of the ability to use MC test-taking cues. The mean CITC values are almost identical between MC and SuperDOMC items. When options are easy to compare, as they are for the SuperDOMC items in this study, the MC items take less time to answer. With SuperDOMC items, options still have to be viewed one at a time.

Despite the overall finding that the average performance of items as measured by the CITC is not different between the two formats, there are differences for the p-value and item latency. However, when looking at individual items, their p-values, CITCs, and response times, certain similarities and differences are apparent. The next three figures—Figures 9.4, 9.5, and 9.6—show these comparisons visually. The items are presented in no particular order.

## Difficulty

Figure 9.4 compares the difficulty (p-values) for each item for both formats, MC and SuperDOMC. Is one format more difficult than another for a particular item? It is expected that idiosyncratic differences might interact with item format.

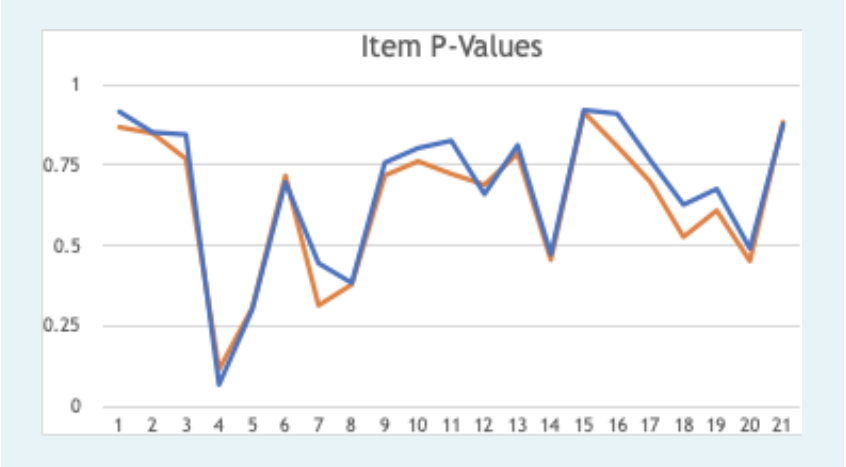



Figure 9.4
Comparison of difficulty for each item.

The patterns seen in Figure 9.4 are reflected statistically in correlations between them. The correlation in item difficulty between MC and SuperDOMC items is very high, almost a perfect positive correlation (r = .98). This indicates that despite the great difference in variability between them, overall item difficulty is unaffected by those variability differences. However, the difficulty is affected by the format (see Figure 9.3). DOMC items remove the ability to use testwiseness and end up being slightly more difficult than traditional MC items.

## Discrimination

Figure 9.5 compares the discrimination values (correct item total correlation, or CITC) for each item for both MC and SuperDOMC. Is one format more discriminating than another for a particular item?

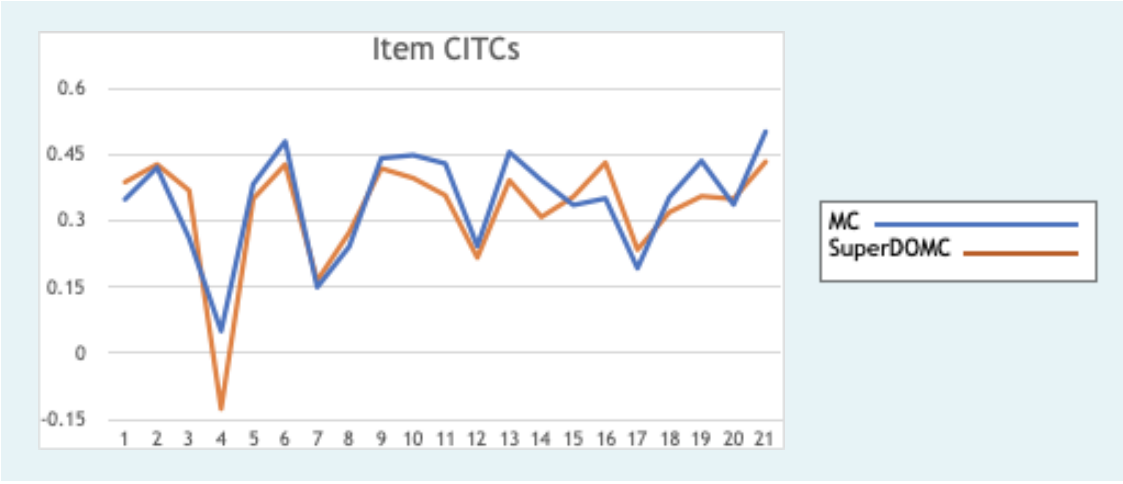



Figure 9.5
Comparison of CITC for each item.

Figure 9.5 shows the close association between the correlations of the MC and SuperDOMC items. In fact, the correlation computed between them is high, at .87. A strong or weak ability to discriminate in one format is quite similar in the other. However, there are differences between items, as SuperDOMC is higher in some and lower in others, compared to MC. Comparing the content of individual items reveals the important reasons for those differences.



As was seen with the p-value comparisons, it doesn't appear that the random variability differences between the item formats have affected overall discrimination values. As Figure 9.3 indicates, the difference between the mean correlation values is only .01.

## Response Time

Figure 9.6 compares the duration (as measured by average response time or latency) for each item in the two formats. Does one format require a response that takes longer than another for a particular item? Given how items are created, it is expected that idiosyncratic differences in response time might interact with item format, whether they are SuperDOMC or not. Scanning the figure, except for one item (#12), the MC versions of the items take less time to answer. Is this difference due to the additional variability introduced by the SuperDOMC item type, or due to some other reason?

The correlation between mean latencies is a very high .96, indicating that it is possible to closely predict the latency of one item format from knowing the other. This ability to predict rules out an effect caused by the addition of random variability in presenting options. Instead, it is probably more productive to look at the item format differences of DOMC versus MC.

When an item's options are simple and easy to compare (as is the case with most of these *Harry Potter* items), the MC version takes less time to answer than the DOMC version. The DOMC format requires the sequential view and responding of individual items, and this process takes more time. When items are wordier or more complicated, as is the case with Item #12 in Figure 9.6, that difference goes away and is often reversed.

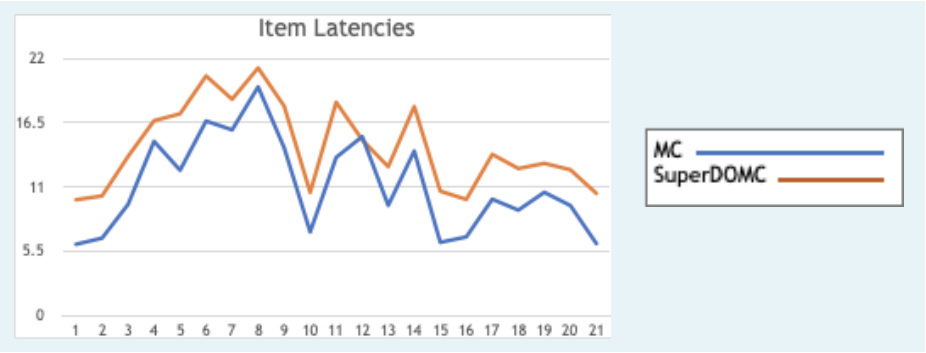



Figure 9.6
Comparison of item duration for each item (shown ranked by difficulty).

## Experiment Discussion and Conclusions

The various analyses in this experiment indicate the MC and DOMC formats perform well within the standards, guidelines, and rules-of-thumb used by psychometricians when evaluating item and test performance. There seems to be no reason to accept or reject MC over DOMC, or vice versa, on the basis of these analyses. The random variability introduced by SuperDOMC items in the content of options does not negatively affect either reliability or validity.

When looking item-by-item, the idiosyncrasies noted in Figures 9.4, 9.5, and 9.6 would suggest item format interacts with specific content. This convinces us that further improvements to our item writing processes can manage difficulty, enhance CITCs, and better understand and predict response time.

A reasonable conclusion from this experiment is that a large amount of randomization, appropriately applied as part of the SuperDOMC, can result in strong psychometric performance on exams.

SuperDOMC items greatly improve the fairness of exams by removing the influence of testwiseness, a long-standing bane of MC-based tests, while not hurting the psychometric quality of an exam.



# Scientific Experiment 3
## DOMC Items Compared to MC Items

**AUTHOR'S NOTE:**

This scientific experiment was designed and conducted similarly to the previous two experiments; much of the description is repeated in the others.

## Introduction

The DOMC item and its extension, the SuperDOMC item, are a step more complicated than the typical MC item. The simplest version of DOMC is to use the exact same content of an MC item, including the same options. All that changes is the format, with each option being presented sequentially rather than at the same time. This difference in presentation:

1. Enhances security
2. Removes the unfairness of testwiseness use.

Referring to the security benefit, an experiment by Gorney and Wollack (2019, personal communication) provided "coaches" with experience on a test comprised of both MC and DOMC items. The coaches were allowed to disclose the test content to "beneficiaries" who could then use the pre-knowledge to get higher scores when they took their tests. The design also restricted which items could be shared, allowing the comparison between "secure" and "compromised" items. The researchers reported these two findings: for secure (non-compromised) items, the beneficiary scores were no different for DOMC and MC items. However, for the compromised items, the scores on the MC items were significantly higher than the scores on the DOMC items. Beneficiaries scored significantly higher on MC compromised items than on MC secure items. In contrast, beneficiary scores for DOMC items showed no difference between compromised and secure items.

For a description of experiments demonstrating the benefits of DOMC to remove effects of testwiseness, the reader is referred to the research of Sonja Willing, Martin Papenberg, Jochen Musch, and their colleagues at the University of Dusseldorf.

DOMC items do not render in as many ways as SuperDOMC or SmartItems because they are restricted to the typical four or five options in the option pool. If an item has five options, an MC version of the item would display all of the five options at once. They could be randomly reordered to display 120 renderings of the item.

On the other hand, a five-option DOMC version of the item with one correct and four incorrect options renders in 193 different ways. Those additional renderings, combined with the fact DOMC items usually display fewer than all five options, is a benefit of the DOMC item type.

Test designers committed to absolute standardization might argue that even this small difference in how the DOMC item is presented is enough to create extra and unnecessary variance in the analyses needed to evaluate items and test performance. This study is intended to test that assumption.

To understand the psychometric properties of DOMC items, it is important to choose a content area and a set of participants where cheating is unlikely. For that reason, the test used in this experiment has almost no stakes for the participants. There were no decisions based on the test, and personal identifying information was not collected. The test was given to individuals from online fan forums. The "fans" were asked to take the test without cheating, a request which a large majority obeyed. The few who used some sort of assistance or outside resource indicated so in a survey question. Their data were removed.

The content of the test was trivia from the *Harry Potter* books by J. K. Rowling. This content was chosen because it allowed the use of a number of large fan forums. The content of the items consisted of:

1. A question asking the research subjects to recall facts from all six books, and
2. A similar number of questions asking the research subjects to make inferences from their knowledge of all six books.

This experiment compared the psychometric performance of DOMC items with their MC counterpart. Both item types were identical in the content for the items. The only difference was in the range and the average number of options displayed for each item. The random error introduced by the DOMC items is in the form of different numbers and sets of options seen by participants, not in the content of the items themselves.

Details about the design of the experiment, the participants, and the test are covered in the Method section below.

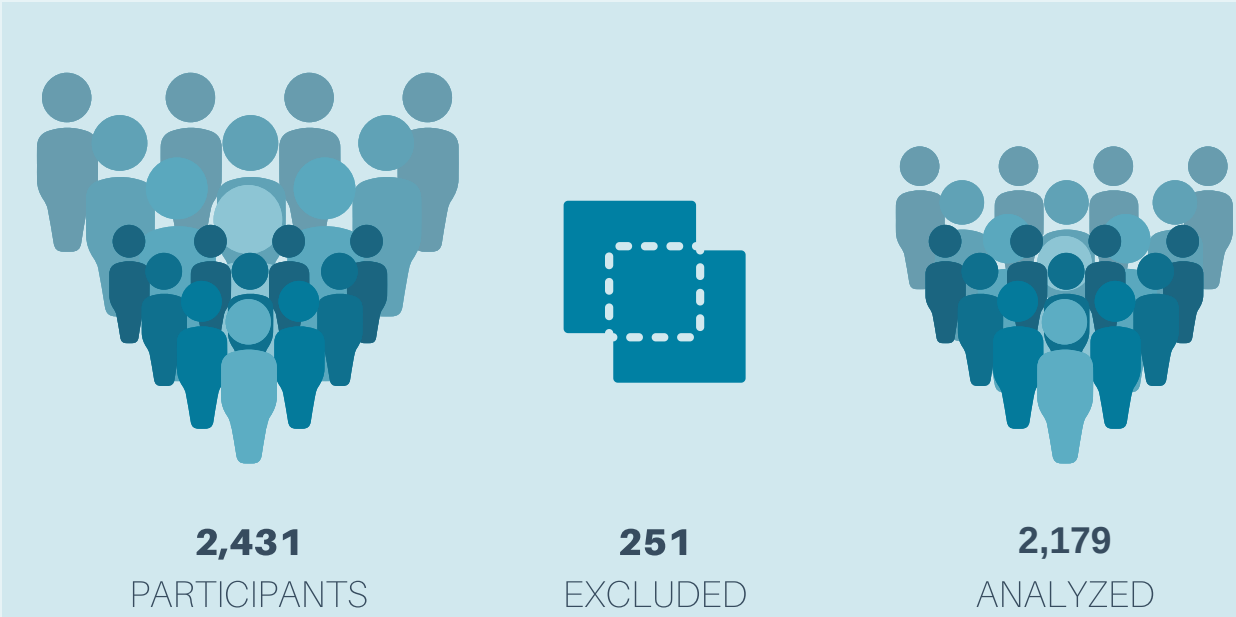
## Method

### Subjects

Participants were volunteers from several online *Harry Potter* fan forums. The subjects were recruited through an online post that included a request for help on a research project and a link to the exam. With a process similar to that used to recruit participants for the other two experiments, the *Harry Potter* forum members were offered a chance to assess their knowledge of the *Harry Potter* books. If they voluntarily chose to participate, they simply clicked on this link:

https://sei.caveon.com/launchpad/harry-potter-by-number

2,431 individuals participated in the study. 251 of the participants had their data excluded for various reasons, including admitting to getting help to answer the questions, or for taking the test more than once. 2,179 individuals remained. These were all first-time test takers who reported they did not receive help answering the questions. The data from those 2,179 participants were used in the analyses.




## Item Design

The 21 DOMC items were created by a *Harry Potter* subject matter expert. They were not created to correspond to particular objectives, so there is no way to classify them as SmartItems. That said, they share important points in common, such as rendering randomly in more ways than their MC counterparts.

Sample DOMC (and MC) Item:

In all seven books, is this the number of times Harry Potter takes Polyjuice Potion?

Options:
0. 4
1. 2
2. 3
3. 5
4. 6

This DOMC item can render in 193 ways, considering the option pool consists of one correct answer and four incorrect ones. The MC item content was identical to the DOMC item.

## Test Design

The experimental design counterbalanced the two item formats across two forms: Form A and Form B. Figure 9.7 shows these conditions.

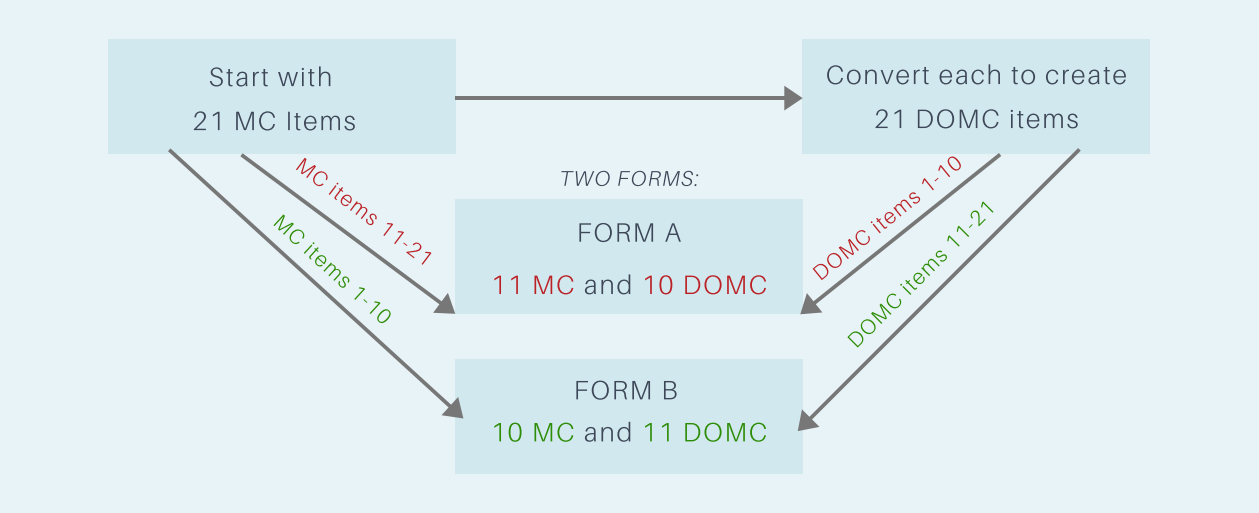



**Figure 9.7**

Randomization was used in four ways in this experiment:

1. The forms were randomly assigned to each participant who volunteered to complete the exam.
2. The 21 items within each form were randomly presented to each participant.
3. The options for both varieties of items (DOMC and MC) were randomly presented.
4. The 21 items of each format were randomly assigned to forms.

The counterbalanced design and the use of randomization to control for known and unknown extraneous variables allowed the direct, unconfounded comparison of formats. The dependent measures used in the study were the responses made by the participants and the latencies for those responses.

## Survey Questions

An eight-question survey was given at the end of the exam. The first three questions asked the participants for their age, gender, and race/ethnicity. The next three asked about the format of the questions, MC or DOMC. The seventh question asked them to rate their expertise relative to *Harry Potter* knowledge. The eighth survey question asked how many times they had taken the test.

Answers to the seventh question were used in the analyses described below to provide validity evidence.

If the difference between MC and DOMC remains unclear, please refer to Miller & Foster (2009) or to the website, www.trydomc.com.

## Procedures

Participants were recruited by a post to several online fan forums for *Harry Potter*. The post included a link to take the exam. (The same link provided above, https://sei.caveon.com/launchpad/harry-potter-by-number.) The volunteers were required to complete a typical reCaptcha security task, and were asked to read about the exam and to follow instructions. Participants were told they would see both MC and DOMC items.

Further instructions disclosed the number of questions (21) they would see and be asked to answer, and revealed they would have 30 minutes to complete the exam. They were informed they would not be able to return to previously viewed and answered items. They were also asked not to seek help from other resources or other persons.

Both MC and DOMC items were described in greater detail, along with what to expect with each. This was followed by a request to respond to a "brief survey" at the end of the test.

Finally, participants were reminded their performance was anonymous.

While the participants answered each exam or survey question, their responses were immediately recorded, along with the time it took to make the response (also known as "response time" or "response latency"). When the experiment concluded, more than 2,300 subjects had successfully completed both the exam and the survey. These results were from first-time test takers only. If the participant indicated by responding to Survey Question #9 that they had used outside help, those data were also discarded. The results were then subjected to statistical analysis.

The analyses included:

- Reliability
- Validity
- Item Performance

## Reliability

With the added variability associated with the random renderings of the DOMC items, would the reliability of the test be degraded compared to a test of entirely static items? Cronbach's alpha reliability coefficient for the MC items was 0.74. For the DOMC items, the coefficient value was .72. Despite the variability difference, the reliability was not affected.

## Validity

One source of validity evidence is to compare the scores on a test with an independent criterion. In this case, that criterion variable is an individual's response to Survey Question #7. It states, "Please rate your *Harry Potter* expertise." Participants chose between "Beginner," "Below Average," "Average," "Above Average," and "Expert." With regard to validity, it is logical to expect that those who answered "Expert" would obtain higher mean scores than those who answered they possessed lower levels of expertise. In addition, a Pearson correlation statistic comparing the responses to test scores should be positive. The mean test scores for those five levels of expertise are shown in Figure 9.8 for both the MC and the DOMC items.

The correlations of response to the survey question and test score for MC were 0.44 and 0.40 for DOMC items. Even though DOMC items render randomly in more ways than MC items, validity evidence is strong and equivalent for both item formats.

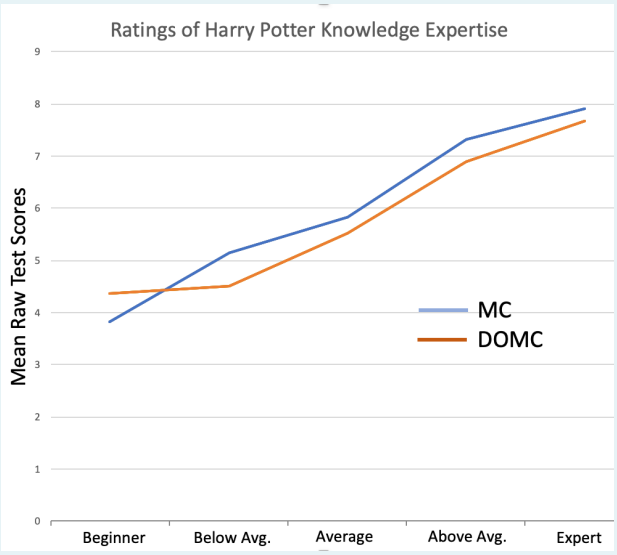



Figure 9.8
Mean test scores for both MC and DOMC questions.

## Item Performance

Most test development efforts eventually result in data being collected from the items administered to test takers. Those data can be analyzed statistically to evaluate each item.

All 42 separate items were analyzed using classical test theory methods, producing estimates of difficulty and discrimination, as well as response time. Difficulty is represented by the p-value, and discrimination by the corrected item total correlation, or CITC.

| Format | P-Value | CITC | Latency |
|---|---|---|---|
| MC | .68 | .31 | 10.78 |
| SuperDOMC | .64 | .30 | 14.06 |

Figure 9.9
Mean items statistics summarized by item format.

DOMC items are usually perceived as more difficult, and that difference is usually a few percentage points as indicated by a lower p-value. DOMC-based items remove most of the ability to use MC test-taking cues (termed "testwiseness"). The mean CITC values are almost identical between MC and DOMC items. When options are easy to compare, as they are for the content of the DOMC items in this study, the MC items take less time to answer. With DOMC items, options still have to be viewed one at a time.

Despite the overall finding that the average performance of items as measured by the CITC is not different between the two formats, there are differences for the p-value and item latency. When looking at individual items, their p-values, CITCs, and response times, certain similarities and differences are apparent. The next three figures—Figures 9.10, 9.11, and 9.12—show these comparisons visually. The items are presented in no particular order.

## Difficulty

Figure 9.10 compares the difficulty (p-values) for each item for both formats, MC and DOMC. Is one format more difficult than another for a particular item? It is expected that idiosyncratic differences might interact with item format.

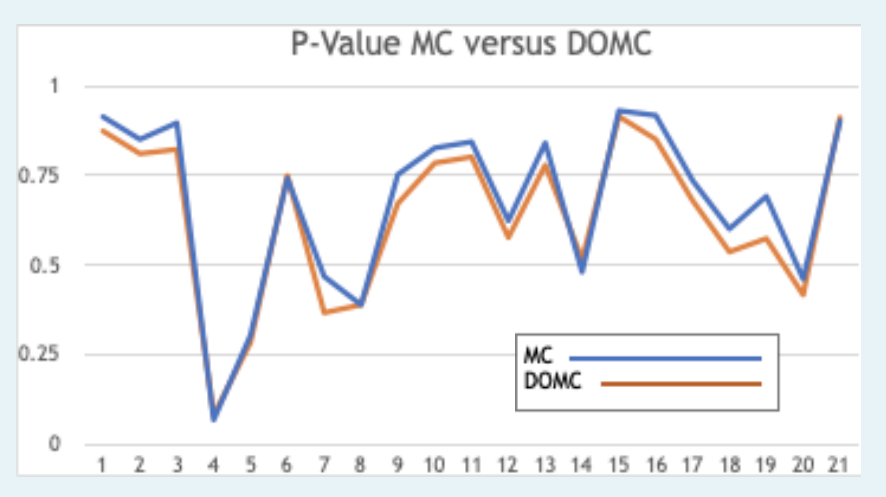



Figure 9.10. Comparison of item difficulty for each item.

The patterns seen in Figure 9.10 are reflected statistically in correlations between them. The correlation in item difficulty between MC and DOMC items is very high, almost a perfect positive correlation (r = .99), indicating that despite the difference in variability between them, item difficulty is unaffected by that variability difference. However, the difficulty is affected by the format. DOMC items remove the ability to use testwiseness and are almost always lower in difficulty by a few percentage points than the traditional MC counterparts.

## Discrimination

Figure 9.11 compares the discrimination values (correct item total correlation, or CITC) for each item for both MC and DOMC. Is one format more discriminating than another for a particular item?

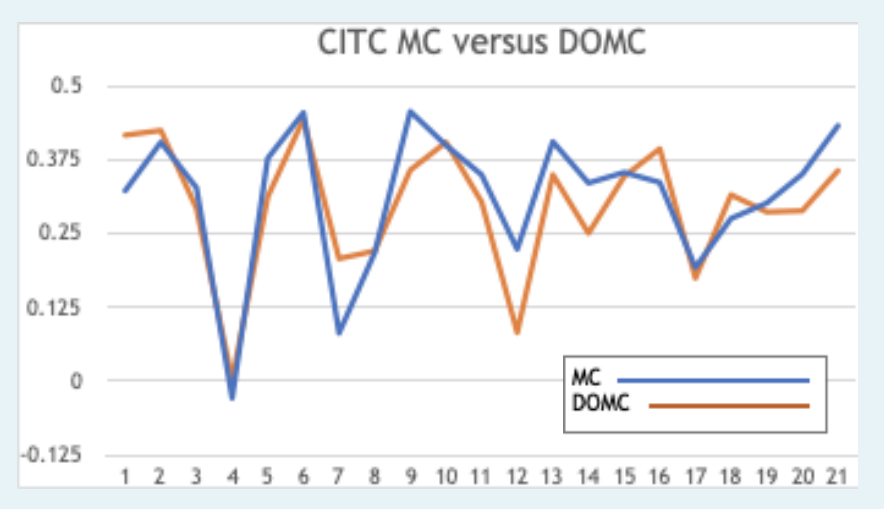



Figure 9.11. Comparison of CITC for each item.

Figure 9.11 shows the close association between the correlations of the MC and DOMC items. In fact, the correlation computed between them is high, at .85. A strong or weak ability to discriminate in one format is quite similar in the other. However, there are differences between items. Comparing the content of individual items will reveal the important reasons for those differences.

As with the p-value comparisons, it does not appear that the large and random variability differences between the item formats have affected overall discrimination values. The difference between the mean correlation values is only .01.

## Response Time

Figure 9.12 compares the duration of the item (as measured by average response time or latency) for each item between the two formats. Does one format require a response that takes longer than another for a particular item? Given how items are created, whether or not they are DOMC, it is expected that idiosyncratic differences in response time might interact with item format.

As the figure illustrates, except for one item (#12), the MC versions of the items take less time to answer. Is this difference due to the additional variability introduced by the DOMC item type, or due to some other reason?

The correlation between mean latencies is a very high .96, indicating it is possible to closely predict the latency of one item format from knowing the other. This ability to predict rules out an effect caused by the addition of random variability in presenting options. Instead, it is probably more productive to look at the item format differences of DOMC versus MC.

When an item's options are simple and easy to compare (as is the case with most of these *Harry Potter* items), the MC version takes less time to answer than the DOMC version. The DOMC format requires the sequential view and responding of individual items, and this process takes more time. When items are wordier or more complicated, as is the case with Item #12 in Figure 9.12, that difference goes away and is often reversed.

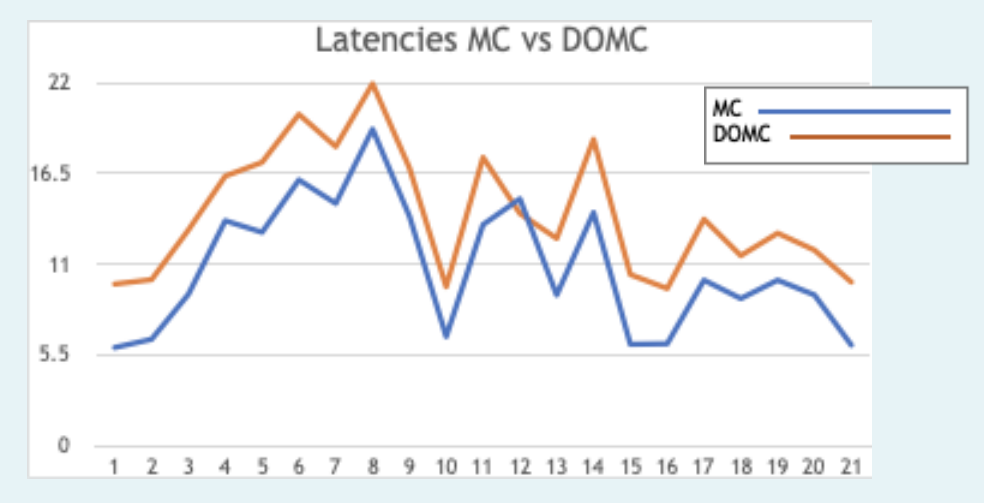



Figure 9.12
Comparison of item duration for each item (shown ranked by difficulty).

## YOU MAY HAVE NOTICED...

The data from Figures 9.9-9.12 are almost identical to the separate study described in Scientific Experiment #2 above. Both experiments used the same content, with Experiment #2 using SuperDOMC questions (with wide variability in the number of available options) and Experiment #3 using the regular DOMC format. The differences in the number of available options had virtually no effect on the item statistics.

# 09 CONCLUSION

The various analyses in this experiment indicate the MC and DOMC formats perform well within the standards, guidelines, and rules-of-thumb used by psychometricians when evaluating item and test performance. There seems to be no reason to accept or reject MC over DOMC, or vice versa, on the basis of these analyses. The random variability introduced by DOMC items in the content of options does not negatively affect either reliability or validity.

When looking item by item, the idiosyncrasies noted in Figures 9.10, 9.11, and 9.12 would suggest item format interacts with specific content. This convinces us that further improvements to our item writing processes can manage difficulty, enhance CITCs, and better understand and predict response time.

A reasonable conclusion from this experiment is that an increased amount of randomization, appropriately applied as part of the DOMC item type, can result in strong psychometric performance on exams. DOMC items greatly improve the fairness of exams by removing the influence of testwiseness, a long-standing bane of MC-based tests, while not harming the psychometric quality of the exam, and simultaneously increasing security.

Chapter

# 10 EVIDENCE FROM DATA SIMULATIONS

## DEMONSTRATING DIFFICULTY EQUIVALENCE

In Chapters One through Six of this book, you learned why SmartItems *should* work. More than that, you learned that we should all *want* SmartItems to work because they would solve some longstanding and important validity issues in the field of testing, specifically the theft of test content, effective cheating, and even testwiseness.

In Chapters Three and Four, we discussed the impact of randomization. We learned that randomization is an effective and popular method that controls systematic error and balances out the differences between items, including difficulty, introduced by SmartItems. We also learned that when using randomization in SmartItems, we are not really doing things much differently than our current day-to-day practices of creating equivalent or equated forms, or pulling items from a pool with CATs or LOFTs. In these chapters, we came to realize we can all accept a bit of randomness in our exams, especially if it eliminates or reduces major sources of systematic error.

In chapters Seven through Nine, we turned our attention to how SmartItems perform. We dove into two case studies where we saw the benefits of SmartItems enacted on two information technology certification exams. The data presented from these case studies indicate SmartItem exams have been functioning properly for the purpose of certifying competence in a field. In Chapters Eight through Ten, you read about carefully controlled experiments. These experiments show SmartItems perform well according to the statistical criteria used to evaluate item performance. It proves tests completely comprised of SmartItems are valid and reliable, at least to a similar degree as traditional tests made up of traditional fixed items.

To expand upon this evidence and provide additional support for the use of SmartItems, this chapter will discuss a data simulation. This simulation has two goals:

1. To demonstrate the equivalence of difficulty for tests with and without SmartItems
2. To show the effect of SmartItems on predicted cheating.

## The Value of Data Simulations

The benefits of data simulations are twofold. First, they are convenient to use, and second, conditions difficult to replicate or impossible to obtain in case studies can easily be replicated.

The simulation described in this chapter assumes that items are pulled randomly from an entire domain (e.g., mathematics skills for an eight-year-old student). This is a case of randomization with no constraints. Therefore, the simulation described uses randomization in a broader way than would be expected if SmartItems were used. On a test of SmartItems, randomization is limited by the design and functionality of each individual SmartItem. In this simulation, there are no such constraints.

Figure 10.0 shows a shape that represents a domain. The dots within the shape represent items. For the simulation described here, the items are selected randomly to serve on a test as illustrated by Figure 10.1.

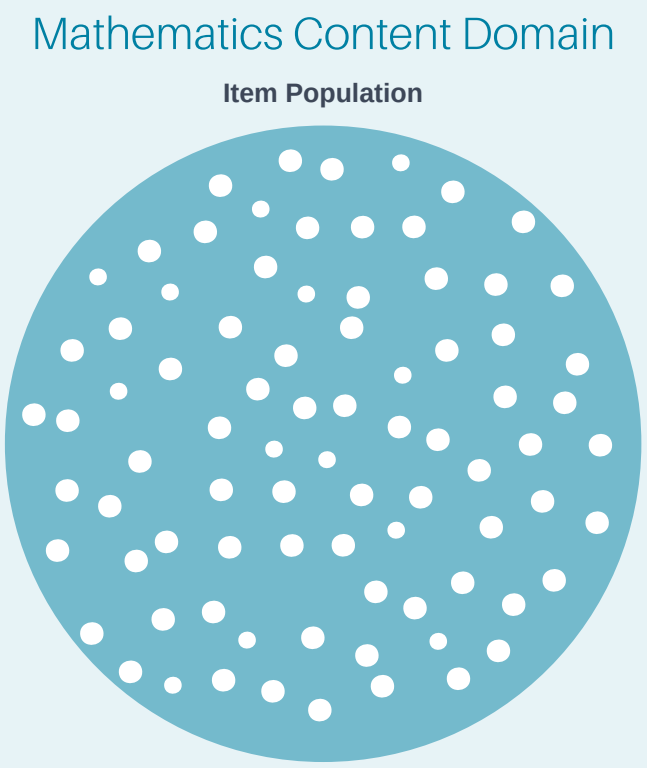



**Figure 10.0**
The circle represents a skill domain (for example, an eight-year-old student's skill with third grade mathematics). Each dot represents an item. Following the example, items are each created to measure any one of the many subskills within the third grade mathematics domain.

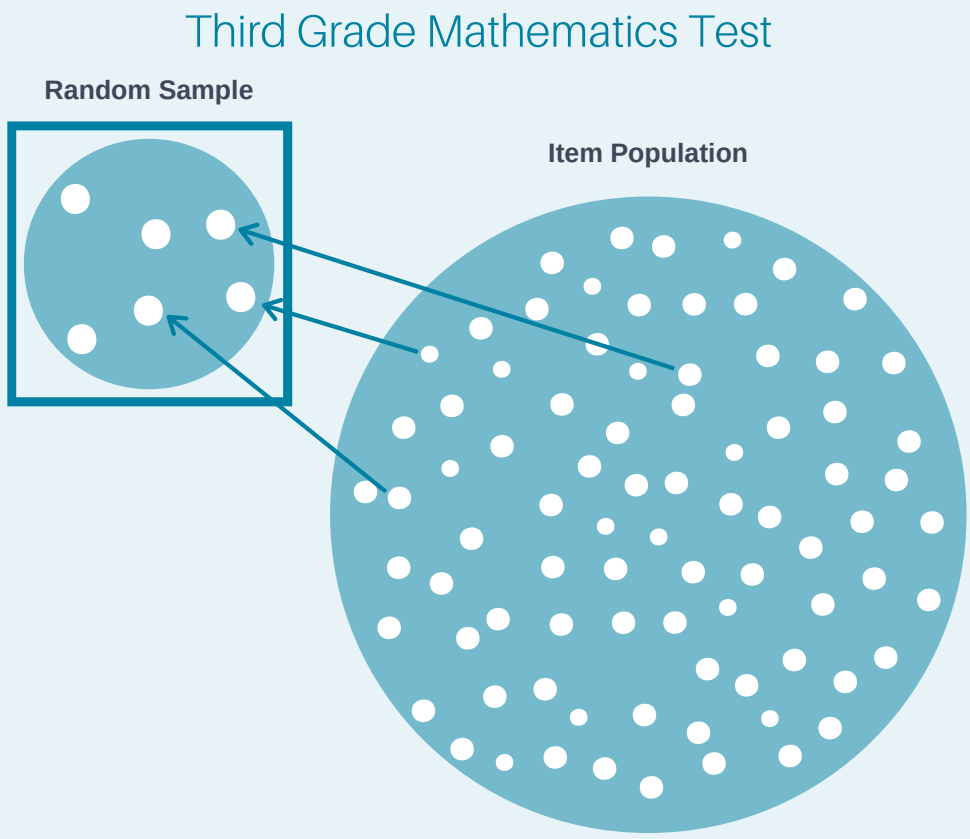



**Figure 10.1**
With the purely random selection process, any item representing any skill within the domain can be selected for a test.

## Simulation: Broad Randomization from a Content Domain

Theoretically, the randomization of SmartItem content introduces error into a test score—this error is random error. With enough items on a test, the sum of that random error should result in a value (slightly positive or slightly negative) close to zero by the end of the test. This theoretical principle supports this conclusion: each score will be minimally impacted by random error, and can be fairly compared to other scores. In addition, the harmful effects of systematic, secondary, and irrelevant variables such as cheating are prevented from impacting the test scores.

However, the question remains: How many items with randomly rendered content are needed for the random error to achieve a value close to zero for each test taker? Is it the same number of items as a typical test? Is it more? Is it fewer? Can we afford the money and time it would take to administer more items than we have been? The simulation described below, first reported by Toton et al. (2018), is an initial attempt at answering those questions.

## The Simulation:

### Data Generation

A bank of 5,000 items was generated using the two-parameter logistic IRT model to describe the distribution of items. The item difficulties were sampled from a normal distribution with a mean *b* parameter of 0.0 and standard deviation of 1.0.[1] The discrimination parameter was sampled from a normal distribution with a mean *a* parameter of 1.0 and a standard deviation of 0.05. Tests were then generated that varied in length from 10 to 60 randomly selected items from the item pool.

A set of 100,000 simulated examinees "took" each of the 51 tests. The exams were normally distributed on their theta values with a mean of 0.0 and a standard deviation of 1.0[1]. In addition to these 51 forms of different lengths, each of the 100,000 examinees "took" a fixed-form test comprised of the same 40 items, randomly selected from the same pool of 5,000 items.

### Test Differences

It was important to come up with a way to determine if a randomized form (of various lengths) was of different overall difficulty than the 40-item fixed-form test. There are several different methods for determining if a person was given a test that is statistically equivalent or statistically different in difficulty from the fixed test. In this case, we computed the standard error of test difficulty using the calculated p-values (classical test theory calculation; proportion of items answered correctly) for items on the fixed-form test.

This was then used to compute a confidence interval for the test difficulty. The standard error of the mean of test difficulty was computed as the standard deviation divided by the square root of test length. This calculation resulted in a standard error of the mean of .028.

Using the standard error of the mean and applying a 95% confidence interval around the mean of the fixed form, we see the confidence interval for the mean difficulty of the fixed form ranged from .44 to .55.



Now, when a simulated examinee received a score (computed as a p value) from a random form, we compared the score of that test to the confidence interval above. If the score (or difficulty) of that random form fell within the confidence interval, we concluded the difficulty of the random form was no different from the difficulty of the fixed form. Otherwise it was treated as non-equivalent.

Finally, we computed the proportion of tests at each test length from 10 to 60 items that were equivalent or different in terms of difficulty, compared to the fixed form. As you can see in Figure 10.2, when the random test was only 10 items long, .363 (or 36.3%) of simulated test takers had a test that was either harder or easier than the 40-item fixed test.[2] That means that only 63.7% of the tests would be considered of equivalent difficulty. However, when the random test was increased to 40 items long, only 8.6% of people had a test score that was either harder or easier than the fixed-form test, or 91.4% of them being equivalent in difficulty.

As the number of items increased to 59, the probability of one being a test of different difficulty continued to decrease. For example, for a random form of 50 items, the proportion of examinees who had an equivalently difficult exam was 95.5%. For tests with more than 50 items, more simulated test takers continued to receive fewer tests that were not equivalent in difficulty to the fixed 40-item test. This pattern is shown in Figure 10.2.

Percentage of Test Takers Receiving Test of Equivalent Difficulty Compared to Fixed Form

10 Item Test: 63.7%

40 Item Test: 91.4%

59 Item Test: 95.5%



Given the assumptions, values, and methods of the simulation, it seems the random error introduced by items with broadly random content on tests (38 to 59 items) will result in roughly equivalent or score-comparable tests. (Remember, the randomization produced in these simulations is greater than what would be produced by SmartItems.)

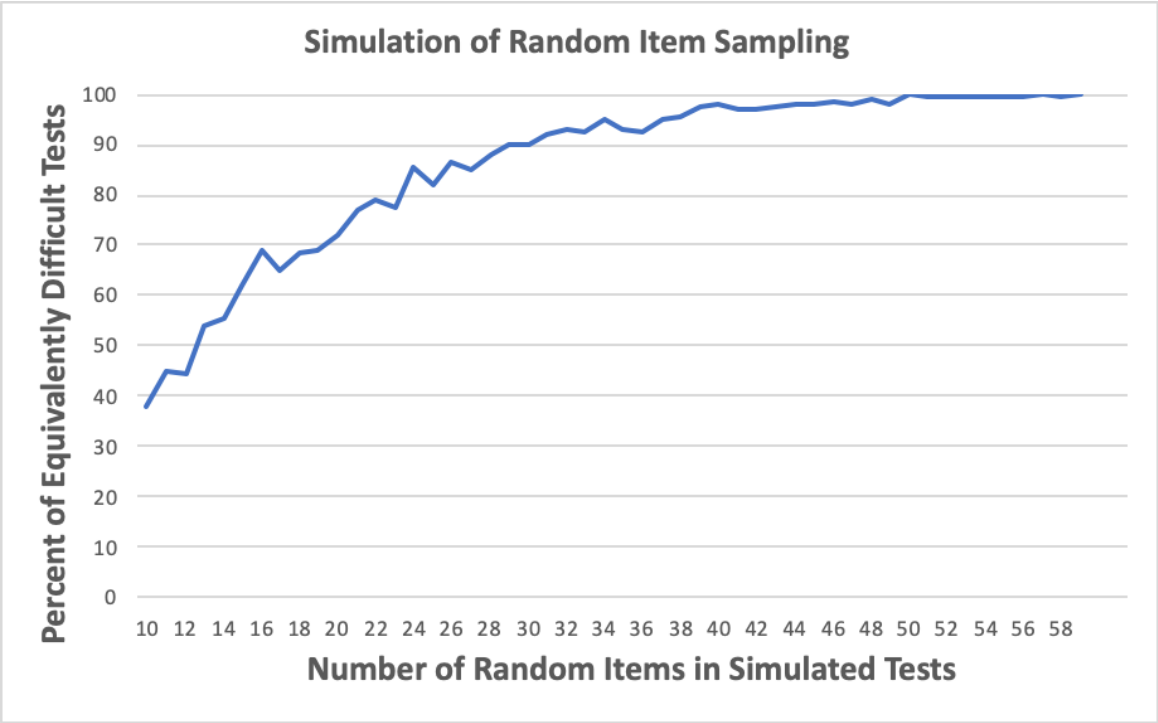



**Figure 10.2**
The results from Simulation #1.

# 10 CONCLUSION

Simulations are a popular method for approximating real testing conditions. The simulation described in this section can be extended by:

1. Simulating the randomization of many individual SmartItems on the same test, and
2. Repeating the simulation, but varying the length of the fixed-length form and changing the distributions of examinee capability and/or item difficulty/discrimination.

The results of this simulation encourage the use of SmartItems when the forms are of sufficient length. If you consider that high-stakes tests used throughout the world for many different purposes (e.g., education, certification, etc.) are usually made up of more than 40 items, then the randomness on tests will have little to no impact on difficulty of the test scores. This is another reason to consider adopting this incredible innovation.

## FOOTNOTES

1. The values for the mean and standard deviation of the ability (or theta) of the examinee sample distribution are common for data simulations.
2. If a test has a mean difficulty that is determined to lie outside of the confidence interval, the test could be easier or more difficult.



CONCLUSION

# THE TIME IS RIGHT

In the book's preface, Dr. David Foster shared his vision of a better world that can be gained through better tests, and realized through SmartItems. I believe in this future, and I hope this book has helped you believe in it, too. If so, it is my heartfelt desire to disrupt the test industry and change it in truly powerful and compelling ways.

The irresistible march of innovation creates impact that permeates every area of our lives. The manner by which we pay bills, plan retirement, arrange travel, navigate new cities, and even relax and enjoy entertainment have undergone sweeping changes that are not only unprecedented, but much improved, too. Time after time, the status quo was challenged. Incredible impacts reverberate through an industry, create new opportunities, solve problems, maximize usefulness, and dramatically reduce costs. While new industries were born, other outdated and staid industries died from a failure to embrace technological change. Technology inevitably transforms—despite institutional reluctance, leaders' trepidation, or a strict

adherence to tradition. SmartItems represent our latest chance to transform testing, and I hope we seize it.

In Chapter One, David detailed the challenges that have hampered our ability to administer and score useful, valid tests. The damage wrought by cheats and thieves is real and undermines not only test score validity, but precious investments and program reputations. The costs—hard dollar costs as well as collateral damage—are staggering.

I've been with Caveon for fifteen years. In that time, I have developed tremendous respect and appreciation for the testing industry. Indeed, I am indebted to it. However, it is reckless to continue to do things in a certain manner just because we always have. If we don't change how we protect tests and test scores from cheats and thieves, I believe the usefulness of our tests will be marginalized to where we, as an industry, simply won't matter much. From where I sit, this type of inertia is unacceptable. History teaches us that if we don't change, someone will change our industry for us. There is evidence that this is occurring every day, but for all our valiant efforts and progress, our industry deserves to flourish and thrive.

SmartItems are a game-changer. For the first time, test scores can be truly protected from the malfeasance of test fraud and the unfairness of testwiseness. Because we finally have the tools to confront these vexing challenges, it is our responsibility—our duty—to do so. The time is right to transform testing and achieve unparalleled validity to test scores by implementing the power and efficacy of SmartItems!

Steve Addicott
COO, Caveon
January, 2020

# REFERENCES


The American Educational Research Association (AERA), the American Psychological Association (APA), & the National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Anderson, N. H. (1982). Methods of Information Integration Theory. Academic Press: New York.

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. Nursing Science Quarterly, 54(1), 67–70. Sage Journals. https://doi.org/10.1177/089431849600900207

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting Performance Standards: Contemporary Methods. Educational Measurement: Issues and Practice, 23: 31-31. doi:10.1111/j.1745-3992.2004.tb00166.x

Crocker, L. M., & Algina, J. (2006). Introduction to Classical and Modern Test Theory. Mason (OH): Cengage Learning.

Cronbach, L. J. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. Journal of Educational Psychology, 32(7), 533–543.

Embretson, S. E. (n.d.). Generating Items During Testing: Psychometric Issues and Models. Retrieved from https://link.springer.com/article/10.1007/BF02294564.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational Measurement (pp. 105-146). New York: Macmillan.

Foster, D. F., Foster Green, A., Thomas, S., & Sowles, J. (2017). *DOMC vs. MC: Summer Research*. Paper presented at the Conference on Test Security, University of Wisconsin, Madison, Wisconsin.

Foster, D., & Miller Jr, H. L. (2009). A New Format for Multiple-Choice Testing: Discrete-Option Multiple-Choice. Results from Early Studies 1. Psychological Test and Assessment Modeling, 51(4), 355.

Gierl, M. J., & Lai, H. (2013). Instructional Topics in Educational Measurement (ITEMS) Module: Using Automated Processes to Generate Test Items. Educational Measurement: Issues and Practice, 32: 36-50. doi:10.1111/emip.12018

Haladyna, T. M., & Downing, S. M. (2004, March). Construct-irrelevant variance in high-stakes testing. Educational Measurement: Issues and Practice, 23(1),17–27. Wiley Online Library. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Kane, M. (2010). Validity and Fairness. Language Testing, 27(2), 177–182. https://doi.org/10.1177/0265532209349467

Kane, M. (2011). The Errors of Our Ways. Journal of Educational Measurement, 48: 12-30. doi:10.1111/j.1745-3984.2010.00128.x

Loe, B.S.  Aiden, D. Condon, & F. Smart, (2019). lisy: Linear Syllogism Generator in AIG: Automatic Item Generator. Retrieved from https://rdrr.io/cran/AIG/man/lisy.html.

Loe, B. S., Condon, D., & Smart, F. (2018). Package 'AIG' [PDF]. Retrieved from https://cran.r-project.org/web/packages/AIG/AIG.pdf.

Lumen Learning. (n.d.). "What is Perception?" *Introduction to psychology*. Retrieved October 10, 2018, from https://courses.lumenlearning.com/msstate-waymaker-psychology/chapter/reading-what-is-perception/.

Matrix Education. (n.d.). Systematic vs Random Errors in Physics: Part 3 of Physics Skills Guide. Retrieved from https://www.matrix.edu.au/the-beginners-guide-to-physics-practical-skills/physics-practical-skills-part-3-systematic-vs-random-errors/.

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. Educational Researcher, 18(2), 5–11. Retrieved from https://doi.org/10.3102/0013189X018002005.

Nunnally, J.C. and Bernstein, I.H. (1994) The Assessment of Reliability. Psychometric Theory, 3, 248-292

Papenberg, M., Willing, S., & Musch, J. (2017). Sequentially Presented Response Options Prevent the Use of Testwiseness Cues in Multiple-Choice Testing. Psychological Test and Assessment Modeling, 59(2), 245–266.

Robinson, P. W. (1981). *Fundamentals of Experimental Psychology* (2nd ed.). Published by Prentice-Hall, Inc. Englewood Cliffs, NJ.

Sireci, S. G. and Zenisky, A. L. (2016). "Computerized innovative item formats," in Handbook of Test Development, eds S. Lane, M. R. Raymond, and T. M. Haladyna (New York, NY: Routledge), 315–334. Retrieved from https://books.google.com/books?hl=en&lr=&id=ed-NAgAAQBAJ&oi=fnd&pg=PA329&ots=Fs-m8RW1Yo&sig=jlKaH4blfYa2DjGwBVJ4T-msI_4#v=onepage&q&f=false

Stanley, J.C.,in R. L. Thorndike (Ed.), (1971). Educational measurement (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

Stecklow, S. (2016, June 10). ACT cancels entrance exam in South Korea, Hong Kong after test leak. Reuters. Retrieved from https://www.reuters.com/article/us-usa-college-cheating/act-cancels-entrance-exam-in-south-korea-hong-kong-after-test-leak-idUSKCN0YX008

Texas Education Agency. (n.d.). Texas Essential Knowledge and Skills for Health Education: Middle School. Retrieved November 13, 2018, from http://ritter.tea.state.tx.us/rules/tac/chapter115/ch115b.html.

Thorndike, R. I. (1951). Chapter VIII: Tests as Research Instruments. Review of Educational Research, 21(5), 450–462. Retrieved from https://doi.org/10.3102/00346543021005450.

Thum, Y. M. (2003). Setting Performance Standards: Concepts, Methods, and Perspectives. Applied Psychological Measurement, 27(4), 305–307. Retrieved from https://doi.org/10.1177/0146621603027004008

Toton, S., Williams, T., Foster, C., Foster, D. F., & Foster Green, A. (2018). *SmartItems and Their Surprising Effects on Test Security*. Paper presented at the Conference on Test Security, Park City, Utah.

Traub, R. E., & Hambleton, R. K. (1972). The Effect of Scoring Instructions and Degree of Speededness on the Validity and Reliability of Multiple-Choice Tests1. Educational and Psychological Measurement, 32(3), 737–758. https://doi.org/10.1177/001316447203200313

University of California, Irvine. (n.d.). Analysis of Errors. Retrieved from http://faculty.sites.uci.edu/chem1l/files/2013/11/RDGerroranal.pdf.

Wikipedia contributors. (2018). Computerized adaptive testing. In Wikipedia, The Free Encyclopedia. Retrieved 21:23, November 13, 2018, from https://en.wikipedia.org/w/index.php?title=Computerized_adaptive_testing&oldid=916096649

Wikipedia contributors. (2018). QTI. In Wikipedia, The Free Encyclopedia. Retrieved 23:47, November 13, 2018, from https://en.wikipedia.org/w/index.php?title=QTI&oldid=866268177

Wikipedia contributors. (2018). Standard-setting study. In Wikipedia, The Free Encyclopedia. Retrieved 23:17, November 14, 2018, from https://en.wikipedia.org/w/index.php?title=Standard-setting_study&oldid=907525235

Wikipedia contributors. (2018). Observational error. In Wikipedia, The Free Encyclopedia. Retrieved 22:02, November 13, 2019, from https://en.wikipedia.org/w/index.php?title=Observational_error&oldid=924999332

Williams, T. (2018). Notes from the Field: SmartItems. The Lockbox, (The Excellence Lockbox). doi: https://thelockbox.caveon.com/notes-from-the-field-smartitems

Willing, S., Ostapczuk, M. & Musch, J. Adv in Health Sci Educ (2015) 20: 247. Retrieved from https://doi.org/10.1007/s10459-014-9528-2

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives* (pp. 19–51). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

# ACKNOWLEDGEMENTS

The process of writing and publishing a book is not easy and would not have been possible without the support of a team of individuals who rolled up their sleeves and tackled this project with a potent combination of optimism, grit, and dedication.

I'm so grateful to the members of Caveon's marketing team, Alison Foster Green, Allie McCabe, and Brinnlie Knavel. Without their countless hours of creative and careful work as well as their gentle promptings for me to do my part, these words never would have evolved from draft to book.

I would like to thank my long-term colleague and friend at Caveon, John Fremer for reviewing this book in its entirety and providing insightful comments. I would also like to thank my colleagues Dennis Maynes, Sarah Toton, Chris Foster, and Tara Williams for their assistance both conducting this groundbreaking research on SmartItems and reading and reviewing sections of this book.

Steve Addicott, our beloved COO at Caveon, read the final draft, and his queries, questions, and suggestions helped this book reach a new level. His unbridled enthusiasm about test security, and SmartItems in particular, is infectious. He sees what the future could be. I am grateful his support as well as his time in writing the conclusion to this book.

A heartfelt thank you to all of you who, both within Caveon and the wider testing community, have helped advanced the idea of SmartItems. Whether through outright support or friendly arguments, you have contributed to this book and to the future of our industry. For all of this, I'm forever and happily indebted.

# RESOURCES

**SmartItems Resources**

- SmartItems Booklet: https://caveon.readz.com/smartitems-booklet
- SmartItems Website: https://smartitems.caveon.com/

**Examples of SmartItems**

K-12 Education SmartItems

- Common Core ELA: examples.caveon.com/common-core-ela
- Common Core Math: examples.caveon.com/common-core-math
- Question on the Periodic Table: examples.caveon.com/periodic-table

Higher Education SmartItems

- Research Methods: examples.caveon.com/research-methods
- U.S. Constitution Bill of Rights: examples.caveon.com/us-constitution

Certification SmartItems

- Maintenance of Medical Certification: examples.caveon.com/medical
- IT Certification—Docker: examples.caveon.com/docker
- IT Certification—Oracle: examples.caveon.com/oracle

Other SmartItems

- National Anthems: examples.caveon.com/national-anthems
- Weather Map Interpretation: examples.caveon.com/weather-map
- Bones Identification: examples.caveon.com/bones

**Information on Discrete Option Multiple Choice (DOMC)**

- About DOMC: http://trydomc.com/about
- Sample DOMC Tests: http://trydomc.com/home
- White Paper "The Discrete Option Multiple Choice Test Item": https://caveon.com/2016/08/19/the-discrete-option-multiple-choice-test-item/

**Test Security Resources**

- Webinars by Caveon (topics include Risk Assessment & Test Security Best Practices, Data Forensics, Investigations, Exam Development, Web Patrol, Site Monitoring, K-12 Education, Technology, & more): https://caveon.com/resources/webinars/
- White Papers by Caveon: https://caveon.com/resources/white-papers/
- The Lockbox test security e-zine: https://caveon.com/resources/the-lockbox/
- Test Security Threats Risk Assessment Tool by Caveon: https://caveon.com/risk-assessment/

**Test Security Conferences**

To see a comprehensive list of conferences, go to: https://caveon.com/events/categories/conference/

# GLOSSARY

### ANOVA

A statistical procedure termed "analysis of variance," appropriate for scientific experiments that are designed to evaluate hypotheses about one or more independent variables. The two-way ANOVA is used specifically for experimental designs with two independent variables.

### Application Programming Interface (API)

A tool to connect software and the computers that use that software so that systems can talk to each other and important tasks can be accomplished without all systems being built similarly.

### Computerized Adaptive Testing (CAT)

A CAT selects items from a large pool, matching the difficulty of an item from the pool to the ability level of the test taker. For example, a high-ability test taker would ideally see only the more difficult items. A CAT is not scored based on the number or percentage of items answered correctly, but on the the level of difficulty of questions he or she is able to answer correctly.

### Confounding Variables

A general term for variables that introduce systematic error into a set of test scores, thereby "confounding" the interpretation of the scores.

### Construct-irrelevant Variance (CIV)

Systematic error introduced into a set of test scores because a factor is being measured along with or instead of the skill that the test intends to measure. A math test that requires a lot of reading may be measuring reading ability along with math ability. Differences in reading ability among test takers would introduce CIV. A variable that raises test scores is referred to as CIV-Easiness, and an example may be testwiseness in a subset of the test takers. A variable that typically lowers test scores is referred to as CIV-Difficulty. Being ill on the day a test is taken would be an example of this type.

### Contrasting Groups Method

This method is used to set a cutscore or standard of performance by comparing the scores of individuals known to be capable or minimally competent and those known to be of lesser capability. By comparing distributions of test scores, a "dividing point" or cutscore is usually easy to discern.

### Correlation

Correlation statistics provide an estimate of the degree to which two variables relate to each other. For example, the heights and weights of a sample of 100 persons are generally positively correlated. That means (continued on next page)



### Correlation (Continued)

that taller persons usually weigh more, and vice versa. A correlation statistic, no matter how significant, is not an indicator of causation, only that the two variables are related. In Chapter 8, the correlation value presented (that is, CITC) refers to the relationship between how test takers answer an item or option, and how they perform on the test as a whole.

### Cronbach's Alpha

Cronbach's alpha, α (or coefficient alpha), was introduced by Lee Cronbach in 1951. It measures the reliability, or the internal consistency, of a test.

### Data Sources

Spreadsheets that contain lists of data. The SME creates these data lists based on content that needs to be included in their SmartItem. The data source could be a list of people or things, or it could be full phrases worded to be correct or incorrect. The data source could also be images, audio files, or video files. An SME can create and use more than one Data Source for a SmartItem.

### Dependent Variable (DV)

In scientific experiments, scientists will measure a variable to see if it changes as a result of the experiment. This is called the dependent variable. The changes in experimental conditions is called the independent variable (IV). In an experiment to see if a particular drug affects cancer cells, the DV might be the growth (or more likely, the lack of growth) in the number of cancer cells before and after the drug treatment.

### Discrete Option Multiple Choice (DOMC)

A selected-response test question format where the options are presented sequentially rather than simultaneously as would be typical in a traditional multiple-choice format. With DOMC, test takers respond with YES or NO, and not all of the options may be presented.

### Error

The influence of factors which are independent of a test taker's True Score.

### Forms

A version of a test. Often, the term used is "test form." For example, for security purposes, a test may consist of one or more equivalent forms. If there are two forms, Form A and Form B, it doesn't matter if a test taker sees one or the other. Scores from both test forms are comparable or can be made comparable by statistical adjustment.

### Graphical User Interface (GUI)

A user interface to a technology that is mostly graphical and is generally considered to be easy to use. Users of most Smartphones and tablets complete tasks on those devices because they encounter GUIs.

### Independent Variable (IV)

In scientific experiments, the independent variable is the variable the scientist manipulates or changes in order to see the effects of those changes on a variable that is simply being measured, called the dependent variable (DV). In a botany experiment, different amounts of fertilizer would represent the IV, while plant growth would be the DV.

### Information Integration Theory (IIT)

A cognitive science theory specifically designed to confirm the accuracy of subjective human judgments. It does this by introducing judges to an integrated cognitive judgment task.

### Item Response Theory (IRT)

A theory of testing that searches for and establishes the relationship between how individuals respond to specific items and their ability on the domain the items/test are designed to measure.

### Linear-on-the-Fly Test (LOFT)

A test design where items are pulled from a large pool during an exam. The items are pulled in a stratified random fashion, constrained by requirements that might include trying to equate difficulty of items on the exams of every test taker. For LOFTs, it is also possible to simply randomly select items from the pool.

### Minimally Qualified Candidate (MQC)

In certification and licensure environments, a minimally qualified candidate is one who should pass a test at a minimal level to be considered capable enough to practice in a profession and to be awarded a credential.

### Observed Score (OS)

The score obtained on a test.

### Psychometrician

A testing professional trained in the design of exams and in the analysis of exam results.

### P-value

The p-value of an item is the proportion of test takers who answer it correctly. It is a measure of difficulty with higher proportions indicating the item is relatively easy and lower proportions indicating the item is more difficult.

### Random Additional Option (RAO)

For a DOMC item, it is possible—when the item has been scored—to provide an additional un-scored option if one is available. This option will appear to the test taker to simply be one more option to answer. Its purpose is to break up any feedback that about the correctness or incorrectness of a response based on whether a new item is presented.

### Random Error (RE)

Error that is assumed to be random, although often it simply represents error that is difficult to control, such as the temperature in the thousands of rooms across the world where and when the SAT is taken. It is easier to consider this to be a random factor than it is to try to make sure every test taker experiences the same temperature. Random error can also refer to error that is introduced into a test by the purposeful use of random process or designs managed by a random number generator or other tool.

### Response Time

Also known as "response latency." This is the amount of time it takes for a test taker to make a response to a specific question or statement. For multiple-choice items, it is usually the time it takes to answer the item. For DOMC items, it is the amount of time it takes to respond YES or NO to specific options that are presented.

### SmartItem API

In 2019, Caveon released the SmartItem API, a tool that makes it possible to create and use SmartItems on any development and test administration engine. Generally, API stands for Application Programming Interface, and allows applications to communicate with one another. The SmartItem API makes it easy for any program (or its vendor) to connect its exam development and delivery systems to interface with it. Once connected, it allows SmartItems to be created and administered within the program's environment, using the program's preferred variety of item types.

### Stem (of the Question)

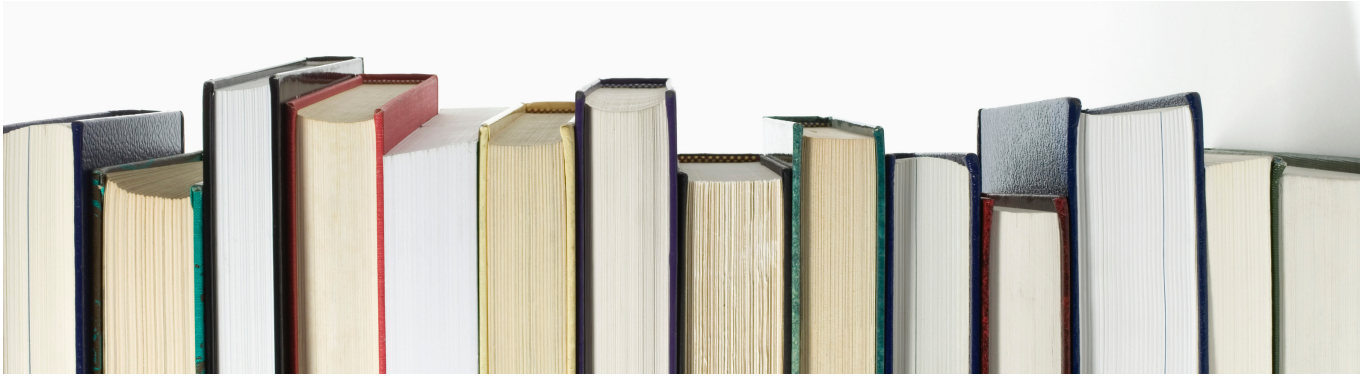
In most item formats, the stem introduces information that is needed to answer the item, and also poses the question or presents the problem that the test takers must answer.

### Stratified Random Sampling

This is a form of randomization with at least one constraint. For example, a research sample may require equal numbers of males and females, which are pulled randomly from pools of males and females.

### Subject Matter Expert (SME)

Individuals within any domain that are most knowledgeable or experienced in the subject matter. SMEs are used to design and author items, to review other items, and to evaluate items so that performance standards (e.g., cutscores) can be determined.



### SuperDOMC Items

DOMC items that go beyond the simple conversion of a 4- or 5-choice multiple-choice item to DOMC. SuperDOMC items may have dozens, hundreds, or even thousands of options. Despite the number of options supporting an item, test takers usually see and respond to only 2 or 3 of them.

### Systematic Error (SE)

Influences on a test score that bias the score so that the score is higher or lower than it should be. Cheating, for example, is a behavior that is intended to increase a test score.

### Systematic Variables

Also known as "extraneous variables" or "secondary variables." These are types of variables that introduce systematic error into a set of test scores.

### Testwiseness

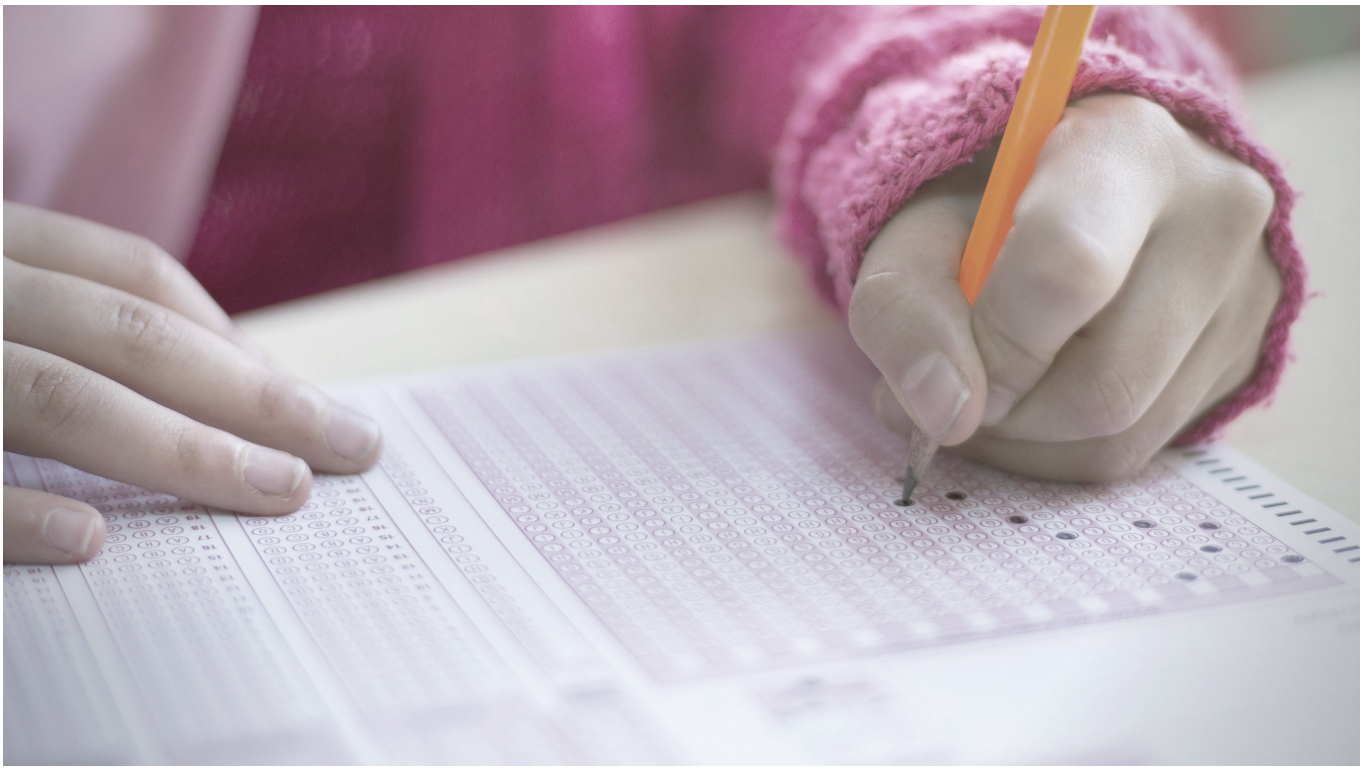
Some test questions, typically traditional multiple-choice formats, provide a large number of cues pointing to the correct option, or pointing away from incorrect options. Some individuals are better at using these cues than others, and achieve higher scores as a result. Those scores are boosted by skills (often referred to as "test-taking skills") that are irrelevant and introduce unfair systematic error.

### True Random Error

Error is introduced in a set of test scores with the help of a computerized random number generator, which then affects a process during testing. For example, the random placement of options for a multiple-choice item would introduce true random error.

### True Score (TS)

The true, but unknown ability of the test taker on the target skills that the test is designed to measure.



© 2020, Caveon, LLC. All rights reserved. Do not copy or distribute without permission.