Exploring Performance Testing in Certification: Lessons Learned and Key Insights from Microsoft

SPECIAL ISSUE

Advancing Beyond Multiple Choice eAssessment

by Liberty Munson and Manfred Straehle

An overview of performance testing and key considerations before adding performance elements to an assessment process. A real-world example is provided as the authors describe why and how Microsoft launched labs in their technical certification program, and lessons learned.

Exploring Performance Testing in Certification: Lessons Learned and Key Insights from Microsoft

SPECIAL ISSUE

Advancing Beyond Multiple Choice eAssessment

by Liberty Munson and Manfred Straehle

When we talk about performance-based assessment, we're talking about assessing people by having them perform tasks or solve problems in an environment that is as authentic as possible. In other words, the test taker demonstrates skills by doing real-world tasks in real-world contexts [1]. Given that broad definition, performance-based assessments can take many forms from having test takers perform tasks in a traditional exam experience (for example, labs embedded in a proctored, secure exam delivery, such as what we do at Microsoft [2]). They may do so separate from a traditional exam experience by submitting a work artifact, work samples, or portfolio (or similar). The assessment could take the form of a practical/hands-on exam (such as the Microsoft Office Specialist Program [3]), OSCE component of the Medical Council of Canada's Qualifying Examination [4], or the National Commission for the Certification of Crane Operators exams [5]. In some cases, a simulated environment is provided for the test taker to complete the tasks or solve problems.

Thinking about Performance-Based Assessment? High Level Considerations

Should you add performance elements to your current assessment program? While you will most likely enjoy higher face validity, fidelity, and authenticity with your test takers [1], these types of elements often require more resources, such as time, staff, and money to develop; and it's easy to over-engineer them [6]. Beyond that, our traditional approaches to psychometrics aren't designed for performance assessments [7]. As a simple example, consider how much time it takes to perform a task, how much time you can ask someone to sit and complete your exam, and you'll quickly realize that you are extremely limited in the number of tasks you can ask someone to perform; this will threaten the validity, reliability, and fairness of the exam. As a result, obtaining accreditation from American National Standards Institute (ANSI) or National Commission for Certifying Agencies (NCCA) may be more difficult when performance elements are included in your certification process.

Design and development. The requirements to develop performance assessments are not fundamentally different from developing more traditional exams, but the process used at each step is likely to be different as you consider each in relation to a performance-based exam. For example, to design and develop your exam, you still need to conduct some form of a job or task analysis, create test specifications and a design document (blueprint), develop items (or tasks, in the case of performance exams), run a pilot or beta test, psychometrically analyze the items, forms, and exams, and set a cut or passing score. Some key differences in these steps will be highlighted in the Microsoft case study, but the notion of pilot testing your performance exam is something worth commenting on in a bit more detail. There will certainly be more cycles in ensuring that the way the performance assessment is delivered works as intended before you even put it in front of test takers to start collecting psychometric data on the tasks themselves. It would be very easy to introduce error into the assessment process if it's not being delivered correctly, the requirements aren't understood, or the test taker has to "learn" something to be able to perform the tasks required. In other words, the beta test is more than just a test of the tasks themselves but also includes some testing of the way the performance elements will be delivered.

Depending on your ability to automate scoring, you will likely also need to consider the following:

- Scaling and equating rubric and rating development
- Raters
- Rater training and calibration
- Inter-rater reliability analysis.

Scoring. Assuming you have decided to add performance elements, you will need to consider how those elements will be incorporated into your assessment process and how they will be scored. Scoring is complicated and will influence how you incorporate elements into your process. This is a bit of a chicken-and-egg situation because how you implement these elements affects scoring and scoring certainly affects how you can implement.

Can, and will, scoring be automated? If so, you'll need to be able to clearly communicate the scoring rules in a way the scoring engine (driven by compute programming) can "understand;" if not, you'll need to design detailed scoring rubrics and train graders (raters, judges, evaluators) to ensure a high level of inter-rater reliability. This implies that more than one human grader should be involved in the scoring process, requiring more resources, such as cost, and increasing the complexity of your current program. Remember

that performance assessments are more authentic evaluations of skills—but only if they are scored correctly, if they are not, they almost always lack validity.

Task inclusion decisions. Next, you need to think about how you will decide which tasks to include in your assessment. As noted, because performing a task generally takes longer than answering a multiple-choice question, you will most likely not be able to include as many tasks as you would like because of seat time limitation. So you need to consider your task sampling strategy or which tasks to include. Task sampling must consider both context (situation/task) and construct (knowledge/skill) dimension. For example, if the purpose of a test is to assess data interpretation skills, it should be possible to develop a list of the types of data that examinees should be able to interpret, a list of the types of situations in which those data are used, and a plan for stratified sampling from those situations, using the frequency and importance of the situations as a guide. Tasks would then consist of a series of high-fidelity simulations of data gathering in those situations.

Now that you have a sense of some of the big considerations when undertaking a performance-based assessment process, let's take a closer look at how performance testing came to life at Microsoft.

Overview of Microsoft's Certification Program and Exams

All Microsoft Certification exams are computer-based, delivered through testing centers or online proctoring, are available around the world, and can be taken at any time. We have a wide variety of item types, including traditional multiple-choice and more interactive item types, such as drag and drop, hot area, and build list. Our goal is to create additional value for our certified candidates and employers by adding performance elements, specifically labs, to each certification. To do that, at a high level, our vision was to design our lab experience such that each lab would contain up to 15 tasks that must be completed in Azure, Microsoft 365, or Dynamics 365, and that in order to complete those tasks, examinees would connect to the technology through the exam interface, ensuring a seamless assessment experience.

Microsoft's Approach to Performance-Based Testing

Microsoft has been dabbling in performance-based testing for more than a decade. We started by converting our Office exams to “in app” experiences in the 1990s. In 2008, we launched simulations on our SQL server exams but learned that we couldn't keep the simulations up-to-date with the rapid pace with which the UI and technology were changing. It was expensive and largely unscalable, especially when we layered in localization. We experimented with emulations in 2010 with our Windows Server exams,

allowing examinees to connect to virtual environments through the internet but found that the user experience in many parts of the world was undesirable, with long lags and load times. In 2018, the story changed... At the time, quite honestly, advances in technology—including increased access to broadband and faster broadband (reducing latency), increased computing power and memory (reducing the time needed to provision the labs, that is, starting parameters needed for the test taker to perform the required tasks), simplification of provisioning requirements (allowing for more flexibility in the starting state of the labs), and so on—were creating an environment where many of the hurdles we faced in launching a global performance-based assessment experience delivered to thousands of examinees daily were reduced.

Introducing Labs

In our case, performance-based testing would take the form of labs. As we considered how to deliver labs globally and at scale, we identified several key goals that were critical to the success of this solution:

- Because our certifications require the application of knowledge and skills to solve problems, we needed to ensure that the labs were provisioned with starting parameters that required the examinee to determine what the problem(s) was and then to fix it.
- Because Azure costs are based on time that the lab is running, meaning costs increase the longer a lab is available, we needed to have a way to dynamically provision the labs at exam launch rather than having them running constantly.
- Azure, Microsoft 365, and Dynamics 365 are browser-based, so the exam would need to be able to open these portals in a secure browser to maximize the fidelity of the experience while still maintaining the security of the exam content.
- Labs needed to be integrated into the overall exam experience and scored automatically through scoring scripts for this to be sustainable.
- Labs needed to reflect the tasks that people in the job role perform in their day-to-day experience. For example, Azure administrators may be asked to add users, set up security groups, or configure a virtual machine. Azure developers may be asked to provision virtual machines, deploy code to a web app, or develop code that uses secrets and certificates stored in Azure Key Vault. Microsoft 365 enterprise administrators might be asked to configure tenancy and subscriptions, create service requests, and configure application access. Dynamics 365 sales functional consultants might be asked to configure sales settings, create accounts, and manage sales order processing. In other words, the solution needed to be flexible

enough to support broad adoption across a wide variety of roles and technologies, allowing us to provision labs in such a way that skills identified in the job task analysis as being critical for success in a given job role could be performed in the lab in a secure testing environment.

- Finally, we know that cloud-based technologies are constantly changing, so finding a way to automate the identification of changes that affected task completion, provisioning, or scoring was critical.

With these goals in mind, we set to work, and in September 2018, Microsoft launched our first lab-based exam.

The final solution ensured that each lab has an environment where independent tasks can be performed by the test taker. Tasks could be performed in any order. The labs are designed based on the time it takes to complete and in some cases to process. Only tasks that can be feasibly performed and completed within the allotted lab time were included. The tasks must have a measurable outcome or “checkpoints” that can be evaluated using scoring scripts. And the labs needed to be designed to support localization/translation.

Approaches to Lab Implementation

We identified three different lab interfaces to connect the examinee to the lab. Figure 1 shows an example of the interface where only one virtual machine is required. As you can see, the Azure portal is on the left and the list of tasks to be completed on the right. This is the most common type of lab implementation as it’s the simplest. It works for job roles that typically perform tasks in one virtual machine.

Figure 1. Lab interface with single virtual machine.

Figures 2 and 3 show examples of a lab interface where multiple virtual machines are required to complete the tasks; this is common for those in Microsoft 365 job roles, such as M365 enterprise administrator and desktop administrator. The examinee can navigate to the tasks and computers by toggling the tabs at the top of the task pane.

Figure 2. Lab interface with multiple virtual machines.

Figure 3. Another example of lab interface with multiple virtual machines.

Finally, some tasks take too long to perform, are overly complicated, require expensive resources, or simply take too long to process. In those cases, we leverage labs as a resource. In these cases, the lab is provided as a resource that the examinee must navigate and explore in order to answer questions related to the lab content. These questions cannot be answered unless the examinee knows how to find the information in the lab (see Figures 4 and 5). These are most common in architect job roles that require understanding a system setup to provide design guidance to solve a particular problem or improve performance.

Figure 4. Example of lab as a resource interface.

The screenshot shows the Microsoft Azure portal interface for Virtual machines. The main pane displays a table of virtual machines under the subscription 'Microsoft AZ-301 4'. The table has columns for Name, Type, Status, Resource group, Location, Source, Maintenance status, and Subscription. Three virtual machines are listed: NVM1, NVM2, and NVM3, all of type 'Virtual machine' and status 'Running', located in 'West Europe' and 'Central US' respectively, sourced from 'Marketplace'.

On the right side, there is a task pane titled 'Task 2' with a note: 'NOTE: To answer this question, sign in to the Azure portal and explore an Azure resource group named ResourceGroup1lod15085424.' Below the note are four multiple-choice options: A. the virtual machine size, B. the public IP address, C. the location, and D. the storage type. At the bottom of the task pane are buttons for 'Help', 'Color scheme', and 'Next'.

Figure 5. Another example of lab as a resource interface.

The screenshot shows the Microsoft Azure portal interface for Storage accounts. The main pane displays a table of storage accounts under the subscription 'Microsoft AZ-301 5'. The table has columns for Name, Type, Kind, Resource group, Location, and Subscription. One storage account is listed: 'store15085425' of type 'Storage account' and kind 'StorageV2', located in 'East US 2' under the resource group 'WebPool-RG-lod15085425'.

On the right side, there is a task pane titled 'Task 4' with a note: 'NOTE: To answer this question, sign in to the Azure portal and explore the Azure resource groups.' Below the note are three multiple-choice options: A. Create a service endpoint that binds to VNET1, B. Configure a service endpoint that allows connections only from 10.10.0.5, and C. Configure a service endpoint that allows connections only from 127.0.0.1. At the bottom of the task pane are buttons for 'Help', 'Color scheme', and 'Next'.

Task Development Key Learnings

Most importantly, not all content works as a lab. Focus on the objectives or skills where it makes sense to build out labs; don't force skills to be assessed by labs. Just because it can be done, doesn't mean it should be done.

Learning from early feedback, we stopped requiring that examinees type in the username and password (they struggled typing them correctly). Rather, they can now log in by clicking the username and password in the task box.

Item writing requires a new skill set. With labs, it is critical that there is collaboration between content subject-matter experts (SMEs) and scripting SMEs. Without this collaboration, the item writer might design an awesome lab that cannot be provisioned, scripted, or scored.

Remember that the goal of labs is to test the examinee's ability to solve problems, so it's key you don't tell them what to do in the task. Tell them the problem and have them solve it.

Finally, tasks should map to real-world behavior in more than one way. Provide some context, but not too much since too much usually leads to trivial information not required to solve the problem. And remember that scoring should not force one method over another. At the end of the day, it doesn't matter how the examinee accomplished the task, only that they did successfully.

Delivery Key Learnings

Integrating labs into the exam experience required significant coordination and commitment between Microsoft, the lab hoster, our exam delivery provider, VUE, and their technology partner, ITS—perhaps much more than anyone really expected. Do not underestimate the challenge of delivering these technology solutions. There will be a lot of technology challenges to solve, so it is important to consider that into your project management plan, especially concerning timelines.

Not all test centers have the infrastructure in place to deliver exams successfully—either because they have outdated equipment or unreliable bandwidth.

Lab exams take longer for examinees to complete, so exam time increases with labs. If you add labs to exams with existing appointments, you will need to cancel or reschedule some appointments.

Finally, if an examinee can remember the login and password, they can access the environment after the exam, so it's crucial that the labs be scored as quickly as possible and that they are torn down immediately. This also reduces costs associated with Azure consumption.

Other Key Learnings

Don't forget about keyboard differences and language considerations if you have an international program. This was especially challenging in countries with non-Roman characters.

You will need to educate your examinees about what to expect. For example, can tasks be performed in any order? Do they need to wait for tasks to execute? It takes time to spin up the labs, which is why we have a section of non-performance items that precede the labs; however, if examinees move through the content quickly, they may have to wait. Along those lines, it takes time to score the labs, so they may have to wait to receive their scores.

When you initially launch exams, you may have more escalations and support requirements. Show empathy and understanding as examinees learn the new experience and you learn what works and what doesn't.

Psychometrically, we are seeing that tasks take three times as long to complete as traditional items [2], they are more difficult, but they have great ability to differentiate high and low performers. On the topic of psychometrics, you will still need to monitor the exam and item performance.

Most importantly, examinees love performance testing as shown from the results from our exam satisfaction survey. Overall satisfaction for exams with labs is approximately 10 points higher than satisfaction with exams without labs. Further, test takers who take lab exams have higher scores across most of the key attributes that we assess in our satisfaction survey (for example, quality of content, degree to which exam assesses real-world skills, alignment with learning content, questions reflect practical approaches to solving problems, to name a few). On average, these attributes are approximately 11 points higher on lab-based exams. This is also illustrated in the comments left by survey respondents:

- "It was hard, really hard—very difficult and that's why this exam meets my expectations. The lab part was something new and incredible for my experience."

- “The interactive labs were resilient and seemed like a real environment instead of just making you navigate to the required task.”
- “The lab exercises are challenging and could apply in real-world scenarios.”
- “Highly appreciated having labs with real-life tasks. Tasks seemed easier than the multiple-choice questions for someone who spends his day in Azure portal.”
- “First Azure lab in exam, enjoyed it.”
- “Questions regarding virtual networking and load balancing were much harder than what I was learning. Labs were much more engaging and interesting than I expected.”
- “The hands-on lab section where you perform skills in Azure is very impressive, technically.”

Higher fidelity, more authentic approaches to assessment are valued by our test takers as evidenced above and proved greater face validity in the assessment process. Moving beyond multiple-choice testing to performance testing is the future of assessments and exams in all areas, including certification, licensure, and education.

Conclusion

Transformation is necessary regardless of industry, and one might say it's long overdue in the assessment industry. Our approach to assessment is rapidly becoming obsolete as the growth of the cloud makes computing and technology more prevalent, substantially increases computing power, and changes the way people interact with the world and their expectations.

Throughout the history of psychometric assessment, we have relied heavily on structured responses, such as multiple-choice questions, on tests because these types of questions are easy to score, making the assessment process scalable, but they are artificial, inauthentic evaluations of skills. Technology, however, is creating opportunities to think differently about the questions we ask, how we ask them, and how we evaluate the responses.

As a testing industry and as testing professionals, we are not delivering on the promise that technology is providing to truly innovate our approach to assessing knowledge, skills, and abilities. While small steps have been taken, our industry is traditionally slow to adopt anything that is truly different and that challenges the status quo, and that is likely to be to our detriment. We need to think big, and even if we can't implement those big ideas, by thinking big we can leap toward innovation in ways that fundamentally change how we approach assessment, and help people not only continue to learn but to demonstrate their abilities in ways that make sense to them.

Prior to COVID-19, there was a rising tide of criticism and skepticism about traditional forms of assessment, and the pandemic was the perfect storm, changing it into a tidal wave, underscoring the smallness of the steps we have taken to leverage technology to change how we assess people. People who have been opposed to traditional forms of assessment saw this as an opportunity to opt out on a grand scale, and the recent decision by California to no longer require the ACT or SATs in their college application process by 2025 is just one example [8].

This highlights the risk that our audience will decide that objective measurement is irrelevant, easily replaced, or doesn't provide sufficient benefit for the associated costs. Further, our reliance on our current item formats, development processes, analytics, test delivery modalities, and psychometrics that have not evolved to accommodate the needs of our students, employees, and other stakeholders, much less today's technologies, big data, and the increasing importance of non-knowledge based skills will undermine the testing industry. To address these risks, we need to understand customer needs deeply and create more appropriate assessments. Multiple-choice questions will not meet this need.

We must rethink our approach to assessment because objective assessment is important. It helps us understand where someone's strengths and weaknesses lie. It helps them learn, grow, and thrive in the constantly changing world we live in today.

Emerging technologies, such as machine learning, artificial and ambient intelligence, gaming, animation, virtual reality, speech/gesture/gaze/voice recognition, blockchain, and bots, just to name a few, can be harnessed to change the world of assessment. Microsoft's inclusion of performance elements on our certification exams is just one example of how testing and assessment programs can begin to leverage the power of technology to provide more authentic skills assessments. We look forward to more testing programs joining us on this journey.

References

- [1] [What is performance testing?](#) The Performance Testing Council. 2020.
- [2] Munson, L. J., Rubin, A., et al. The past, present, and future of technology in performance assessment. Breakout session accepted at Association of Test Publisher's Global Annual Innovations in Testing Virtual Conference, 2020.
- [3] [Certiport](#). Microsoft Office Specialist Program. 2020.
- [4] [OSCE: Definition, purpose and format](#). Medical Council of Canada. 2020.
- [5] [National Commission for the Certification of Crane Operators: Certifications](#). 2020.

- [6] DiCerbo, K. A future of assessment without testing. Keynote presented at the Beyond Multiple Choice conference, 2020.
- [7] DeChamplain, A. [“New” psychometrics](#). A webinar panel discussion presented by Association of Test Publishers. 2020.
- [8] Boyette, C., Silverman, H., and Waldrop, T. [University of California system will no longer require SAT and ACT scores for admission after settlement reached](#). CNN. May 15, 2021.

About the Authors

Liberty Munson is the director of psychometrics for the Microsoft Worldwide Learning organization and is responsible for ensuring that the skills assessments in Microsoft Technical Certification are valid and reliable measures of the content areas that they are intended to measure. She is considered a thought leader in the certification industry, especially in areas related to how technology can fundamentally change our approach to assessment design, development, delivery, and sustainment, and has proposed many innovative ideas related to the future of certification. Prior to Microsoft, she worked at Boeing in their Employee Selection Group, assisted with the development of their internal certification exams, and acted as a co-project manager of Boeing’s Employee Survey. She received her Bachelor of Science in psychology from Iowa State University and master’s degree and doctorate in industrial/organizational psychology with minors in quantitative psychology and human resource management from the University of Illinois at Urbana-Champaign.

Manfred Straehle is one of approximately 20 assessors for the ANSI ISO/IEC 17024:2012 accreditation standards. He possesses immense proficiency as a consulting assessment and educational research expert, and sits at the top of Assessment, Education, and Research Experts (AERE) as founder and president. Dr. Straehle has held roles at Advertising Research Corporation (ARC), Media Broadcasting Company (pharmaceutical marketing), National Board of Medical Examiners (NBME), Genesis Healthcare, Prometric, and Green Building Certification Institute (GBCI), to mention but a few. His long professional history and wealth of experience make Dr. Straehle immensely important to the operational value of AERE. Dr. Straehle is no stranger to leadership responsibilities and has held several leadership positions for the credentialing and standards industry association while serving in the leadership capacity for countless other committees and institutions. A proud member of reputable professional societies such as the American Evaluation Association (AEA), National Council on Measurement in Education (NCME), and Association of Test Publishers (ATP), Dr. Straehle leads AERE with a clear vision of excellence and a team-based collaborative approach.

DOI: 10.1145/3470744