Brigham Young University
BYU ScholarsArchive

Instructional Psychology and Technology Graduate Student Projects

Instructional Psychology and Technology

2023-04-17

A Process for Improving the Quality of Multiple-Choice Certification Exams

Tanner Kohler
tannerkohler@gmail.com

Follow this and additional works at: https://scholarsarchive.byu.edu/ipt_projects

BYU ScholarsArchive Citation

Kohler, T. (2023). A Process for Improving the Quality of Multiple-Choice Certification Exams. Unpublished masters project manuscript, Department of Instructional Psychology and Technology, Brigham Young University, Provo, Utah. Retrieved from https://scholarsarchive.byu.edu/ipt_projects/56

This Evaluation Project is brought to you for free and open access by the Instructional Psychology and Technology at BYU ScholarsArchive. It has been accepted for inclusion in Instructional Psychology and Technology Graduate Student Projects by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

A Process for Improving the Quality of Multiple-Choice Certification Exams

Project and report: Tanner Kohler

Advisor: Dr. Randall Davies

Instructional Psychology and Technology MS program

Brigham Young University

Abstract

A report on the process developed for improving multiple-choice UX certification exams for Nielsen Norman Group. The main problems with the existing exams (poor content validity, too many easy questions, low overall discriminatory power, and poor adherence to multiple-choice question best practices) are described in detail. Measurable goals for addressing these problems are also described. A detailed plan for tackling each individual exam, which can be applied in other contexts, is presented with related resources.

Introduction

Practitioners in many industries participate in professional training to advance their personal knowledge and abilities relevant to completing high-quality work, moving to a better position, or transitioning from one career to another. In for-profit business environments, one major motivation for attaining additional training is gaining recognition as an expert in one's field (Daniels, 2011). In many cases, such recognition is the gateway to increasing one's salary, receiving offers for new job opportunities, being given more authority in one's current role, and completing higher-quality work.

The field of user experience (UX) design is relatively nascent but growing quickly, with the largest proportion of current practitioners entering the industry within the last five years (Nielsen, 2017). This means that many UX practitioners are only beginning to master the knowledge and skills necessary for completing high-quality work or advancing in their careers. Many new UX practitioners feel the need for additional training to supplement their basic knowledge of best practices. For many of these designers, formal UX training comes through professional seminars, workshops, boot camps, or conferences. These opportunities cover basic UX topics and allows them to begin putting UX design knowledge into practice (Getto & Beecher, 2016). There are only a small number of college degrees that are entirely devoted to UX. Additionally, many UX practitioners have received no formal training and have come into their current role out of necessity, not because they have been trained to do it. Thus, much learning occurs on the job.

It is not yet well-defined what "competence" looks like for newly developing UX professionals, making it difficult for hiring managers and organizations to identify qualified candidates with the knowledge and skills necessary to succeed (Gray, 2014). One study demonstrated that more experienced UX professionals considered a broad knowledge of many basic UX concepts to be more valuable in a junior UX team member than a deep mastery of a few skills (Gray, 2016). This priority could reflect the realization that much UX training will come on the job. However, because so many current UX practitioners are "laying the tracks before the train," as they both practice UX and learn about it, they require training that will help them put new concepts and principles into practice by applying them to real UX-related problems (Getto & Beecher, 2016). One experienced UX practitioner put it this way, "I think the methods themselves are quite rudimentary [...] But when it comes to actually getting the right value out of them, it's having that right mindset—what are the right questions we need to ask? How can we answer them? And then using that as the basis for what methods you need" (Gray, 2016, p. 4051). It's not enough to simply know *about* basic UX

methods and concepts; it is also necessary to know *why* they are important and *how* to apply them.

In essence, UX practitioners will be most valuable in a job with a basic ability to recall, understand, and apply basic UX concepts (Bloom, 1956; Krathwohl, 2002). For this reason, many practitioners are highly motivated to achieve some kind of recognizable certification to signify their competency in this new and developing field. Because not all training is of the same quality, it is common for UX practitioners to seek out training from trusted, authoritative organizations in the industry. However, simply attending high-quality training does not equate to mastery of those topics. It is also necessary to assess mastery in order to provide a reputable certification that holds some degree of credibility. While some certification programs provide training content and reward learners with a mere "rubber stamp" that signifies participation more than mastery, the certification provided by programs that truly assess learning and competency is of much greater value.

Being a UX industry leader since 1998, *Nielsen Norman Group* (NN/g) has provided training on various UX-related topics for many years. This includes nearly 15,000 practitioners who have received certification for completing at least five courses and passing the related five multiple-choice certification exams. However, NN/g is no exception to the necessary requirements for providing a reputable accreditation for mastery of UX topics. Their certification exams must be more than a rubber stamp to hold their value and provide practitioners with the boost they are looking for by completing their training.

Credible certification programs rely on sound assessments of what participants have learned. Quality assessments are valid, reliable, and have clear utility. A valid assessment accurately measures what it is intended to measure, and the results are interpreted and applied in appropriate contexts (Miller, Linn, & Gronlund, 2013). This means that if a course teaches learners content at a recall or understanding level (Bloom, 1956; Krathwohl, 2002), the assessment should accurately determine the extent to which learners can recall and explain that content. A valid assessment should also fairly examine the same topics that are covered in the course (content validity). A reliable assessment will show consistent results across multiple test-takers and throughout time, demonstrating that it is a robust measure of learning in varying contexts (Miller, Linn, & Gronlund, 2013). An assessment with good utility will have a clear purpose and be helpful in fulfilling that purpose. For example, certification exams should highlight the course topics that learners understand well and those they do not understand. This feedback can provide clear guidance regarding future iterations of the course to better help learners master the intended learning outcomes. Quality

certification exams should also be appropriately situated to assess cognitive gains. Bloom's Taxonomy for the Cognitive Domain (Bloom, 1956; Krathwohl, 2002) is one of the most commonly-used frameworks for guiding the creation of both learning objectives and appropriate assessment items at certain levels of cognitive difficulty. For example, exam creators can use this framework to create exam items that assess recall if a training only presented content on a definitional level, Given these learning objectives, learners would not be expected to demonstrate higher-level thinking related to that topic. Adhering to a framework in this way can be helpful for creating exam items that are neither too easy nor too difficult for learners given the instruction they have received.

Exam developers commonly rely on a few indicators to determine the quality of an assessment. These indicators include basic item statistics such as the difficulty and discriminatory power of each item. Item difficulty is measured by the percentage of test-takers who answer an item correctly. The higher the percentage, the easier the question. Discriminatory power is a correlation between the number of test-takers choosing the correct answer option, and their performance on the exam overall. The higher the correlation, the stronger the relationship between those who answer the item correctly and those who do well on the exam overall. A high discriminating power indicates that the item was answered by test-takers who seem to understand the topics covered in the course. A low discriminatory power indicates that the item seems to be equally easy for those doing well and those doing poorly overall. Discriminatory power has no meaning when all students get the item correct or get the item wrong. There must be some variance in the results to use discriminating power. These statistics are used to flag items that may need revision. They do not indicate the item is of good quality – this is left to the exam creator to determine.

Description of the Evaluand

NN/g offers courses focused on various topics within the field of UX. Each course is a 6-hour, synchronous, remote offering conducted through two-way video conferencing by one instructor. Courses generally have between 20 and 120 participants who join from over 100 different countries. All courses are offered in English only, requiring attendees to have a working proficiency in English.

Course attendees can purchase the opportunity to take a 30-question, multiple-choice certification exam associated with any course they have completed to prove their mastery of the topics covered. Test-takers have 35 days and three attempts to pass the exam. They must attain a score of 80% or better to receive the credit. Credit is

awarded on a pass/fail basis. Successfully passing the test certifies they have gained adequate knowledge and understanding of the topic.

NN/g offers over 40 different courses at any given time, each of which has its own associated exam. Each exam draws 30 random questions for each test-taker from a bank of roughly 30-50 questions. Each exam question has traditionally been categorized as either a "knowledge" or "practical" question using metadata assignments. Generally, "knowledge" questions have focused on attendees' recall and understanding of basic concepts and principles covered in the course, loosely mapping onto the remember and understand levels of Bloom's Taxonomy (Bloom, 1956; Krathwohl, 2002). "Practical" questions have focused on applying concepts to real-world scenarios, mapping onto the *Apply* level of Bloom's Taxonomy (Bloom, 1956; Krathwohl, 2002). Each exam has been designed to pull a portion of the 30 random questions from the "knowledge" group and a portion from the "practical" group to ensure a mix of question types. The number of questions pulled from each group varies by exam based on how many questions exist within each category in the question bank, but have roughly been half and half.

Because each exam is generally taken by at least one individual on a monthly basis (if not more frequently), there is a large amount of accumulated historical data available to calculate basic item statistics. Fortunately, our exam system (onlinetesting.net) automatically calculates the basic item statistics for each question. The system calculates and provides the item difficulty and discriminatory power for each correct option and the distractors for each question (see figure 1).

19	Item difficulty	Discrimination index	Why is it important to have a clear and defined content design process? Question average score: 88.6% · Weight: 4 · Seen 44 times · Category: Understand · Topics: Content Design Process
a.	0	0	a. It isn't, it's better to be agile, flexible, and iterative about it
b.	**88.6%**	**0.260**	**b. Without a process, everyone can have the same content goal but very different ideas about how to achieve it**
c.	9.1%	-0.132	c. Content is often very individualistic and processes ensure designers follow all of the rules
d.	2.3%	-0.105	d. So that content designers are treated with the same respect as other process-oriented roles, such as engineers

Figure 1: Example Item Statistics

Purposes of the Evaluation

This project had two main purposes: 1) evaluate the current assessment practices of all NN/g exams and 2) create a protocol for assessing the validity of and improving

individual certification exams. NN/g initially implemented its certification program as an experimental offering. At the time, many courses were already being taught and had been taught for many years. Because of the initial experimental nature of the certification project, exams were created post hoc to assess what was being taught in the existing courses. This order of things is contrary to the typical way instructional designers are taught to create courses and the associated exams, however, it is a typical process for untrained instructors functioning as instructional designers. Additionally, the exams were created in a low-fidelity form as a test to see how interested attendees were in the offering rather than as a rigorous assessment of the course content.

As the certification program showed promise, more exams were created to expand the offering. However, this was done with the acknowledgment that the exams may not be valid assessments of a course's intended learning outcomes. The team responsible for the maintenance of all exams has long planned to improve the overall quality and rigor of the course assessments used for certification.

Methods

I began this project by conducting an initial evaluation of a sample of the NN/g exams to identify the main existing problems that needed to be addressed, and the severity of these problems. To evaluate the overall state of all NN/g exams, I took a random sample of nine exams (including roughly 30-50 items per exam, and 395 items total) to serve as a representation of the remaining 30-40 exams. I only included nine exams in the sample because the averages for each metric I analyzed stabilized at that point and did not continue to change with additional data.

To make use of the available item statistics, I began by calculating the percentage of items that were "very easy" on each exam, and on average across all nine exams. Based on standards previously determined by NN/g, I defined "very easy" as any question answered correctly by 95% or more of test-takers. To calculate the average difficulty, I counted the number of "very easy" items for each exam individually and divided the sum by the total number of questions on the exam to reveal what percentage were answered correctly by nearly all test-takers. This told me how easy each individual exam in the sample was. I then calculated the average of the percentages from all nine sample exams to give me an indication of how easy the exams were overall (see Appendix A). To better understand what contributed to the overall ease or difficulty of each exam, I also noted the level of Bloom's Taxonomy at which each question was situated (Bloom, 1956; Krathwohl, 2002).

Additionally, I calculated the average discriminatory power for each of the nine sample exams to show me how well they were able to differentiate between test-takers (see Appendix A). I excluded the discrimination power from the averages for any questions which were answered correctly by 100% of test-takers (n = 52) because items with a discriminatory power of a perfect 1.0 was meaningless for my purposes. A perfect correlation, in this case, indicated that all test-takers who had seen that item answered it correctly, meaning that the item had no ability to discriminate between test-takers whatsoever. Including the discriminatory power of these questions in the averages would have led to an inflated overall estimate.

Lastly, I evaluated the extent to which many of the items in the nine sample exams adhered to best practices for writing multiple-choice exam items (Miller, Linn, & Gronlund, 2013) to identify common issues with the items which would need to be revised.

Results

My evaluation of the nine sample exams revealed that there was room for improvement in many places. The following four problems were identified during the evaluation, which are subsequently described in more detail:

1. Poor content validity
2. Many easy questions
3. Low discriminatory power
4. Poorly-written items

Poor Content Validity

All questions for a given exam were originally constructed by the creator(s) of the associated course. In most cases, these course creators are also the regular instructors assigned to the course — often teaching it on a monthly or bi-monthly basis. Changes or updates to exams tended to occur on a needs basis as course content was changed and refreshed. In most cases, questions that assess topics no longer covered following course updates have simply been removed. Slowly removing questions has reduced the number of available questions in each question bank because very few instructors are interested in creating new questions or updating older, unused questions unless required to do so. This natural attrition of questions has contributed to lower levels of content validity as there has been no formal attempt to ensure fair representation of the course content being assessed on each exam.

Additionally, the original creation of the exams for each course did not use any formal method for ensuring some degree of content validity. Untrained in assessment practices, instructors are typically unaware of tools such as a table of specifications (Chase, 1999) that can be used as test blueprints and ensure that course content is adequately represented throughout the exam. Many instructors reported that they quickly scanned through course slides at the last minute to create exam questions on any topic that seemed easy to test, the "low-hanging fruit."

Many Easy Questions

Averaged across the nine sample exams, I found roughly half of the items were very easy, being answered correctly 95% or more of the time by test-takers (see Appendices A-C). Many more questions fell just below the 95% threshold, indicating that the exams had a vast number of very easy questions. A review of items from the sample exams also revealed that many of the questions resided at the "remember" level of Bloom's Taxonomy (Bloom, 1956; Krathwohl, 2002) which undoubtedly contributed to the overall ease of the exams (see Appendix D). This is likely because it takes more skill and effort to create questions at higher levels of the Taxonomy, and because basic recall-based questions, such as definitions, came to mind most easily for hurried instructors who had no test blueprint from which to work. An overabundance of recall-based questions is not uncommon for these types of certification exams (Alzu'bi, 2014; Muhayimana & Nyirahabimana, 2022). Recall-based questions are particularly easy for students to answer as all exams are open-note, and attendees have access to all the course materials and slides after completing the course.

Another factor likely contributing to the overall ease of many items is the inclusion of distractors that are implausible. In many cases, an item might require a test-taker to demonstrate a deeper understanding of content material, or apply it to a specific context rather than simple recall, but the distractors are not plausible, and the correct option is obvious. Further investigation into the discrimination indices of all distractors could shed more light on this explanation, if necessary.

Low Discriminatory Power

In addition to the average item difficulty calculations, I also calculated the average discriminatory power across the sample of nine exams, excluding those items which were answered correctly 100% of the time. Doing so revealed an average DI of 0.12 across the nine exams, with a range from 0.05 to 0.2 (see Appendix A). Within each exam, there were some items with a strong discriminatory power (0.2 or above), and some items with very low discriminatory power (0.05 or below) which is to be expected

on any exam (see Appendix C). However, an abundance of items with low discriminatory power inevitably results in an exam that cannot meaningfully differentiate between test-takers who have mastered the content and those who have not.

Poorly-Written Items

None of the original exam creators have had much training in best practices for assessment creation. Most were unfamiliar with basic guidelines for writing multiple-choice questions. An initial review of items from various exams revealed that some violated basic best practices for writing multiple-choice items (Miller, Linn, & Gronlund, 2013) (see Appendix E). While having questions that are easy to answer is not inherently problematic, it is problematic to have questions that are easy to answer because they were poorly written. Flawed items neither discriminate between test-takers nor assess overall learning. Some of the most common violations included:

- Not presenting a meaningful problem in the question stem
- Including irrelevant material in item stems (particularly for scenario-based items)
- Unnecessarily long alternatives with superfluous words
- Unequal lengths of alternatives, most often with the correct alternative being the longest
- Verbal clues between the item stem and the correct answer
- Using distractors that are implausible (i.e., obviously incorrect)

It is likely that the violation of multiple-choice question best practices, the high number of recall-based questions, and a large number of implausible distractors contributed to the overall ease of many exams.

Recommendations

Based on the results of the evaluation of the nine sample exams, it became clear that I would need to create a process for more thoroughly evaluating and improving each NN/g exam. The following recommended process could be applied by myself or other members of the NN/g team to any exam, and guide the associated instructors through the process of addressing the four above-stated problems. This process strives to help any exam meet the following standards:

1. **Content validity.** The number of items focused on each course topic should fairly reflect how that content was covered during the course.

2. **Difficulty.** No more than 15% of items on an exam should be very easy (i.e., answered correctly by 95% or more of test-takers). However, no items should be overly difficult (i.e., answered correctly by less than 65% of test-takers).
3. **Discriminatory power.** Around 50% of items should have a discriminatory power at or above 0.2.
4. **Multiple-choice best practices.** All items should be evaluated against guidelines for writing that type of item to avoid any common pitfalls for writing multiple-choice items; making them unintentionally easy or difficult (see Appendix G).

In discussing these goals with relevant stakeholders at NN/g, it became clear that they were lofty and would potentially require many hours of work for each exam, depending on its current state. This led to a secondary goal for the project (which is true of most projects) was to find a way to evaluate and improve each exam as efficiently as possible, minimizing the required time and cost. The process to be created would need to be adaptable to the state of each exam so as to salvage any existing questions and thus reduce the amount of time required on the part of the exams team and pertinent instructors.

Proposed Exam Improvement Process

Through multiple rounds of iterations, we developed a process with four phases for evaluating and planning improvements for each exam: (1) Content Analysis, (2) Gap Identification, (3) Exam Updates, and (4) Implementation. We piloted this process with one, well-established NN/g course that had a large set of data available, providing robust item statistics. It was a course that is taught frequently and would be able to quickly gather fresh data to reveal how well the process was able to achieve the project goals.

Each phase is described below with an outline of the necessary steps in that phase. This section of the report is framed as step-by-step instructions for reverse engineering an exam to compare its current state with a desired future state. Each step has an associated time estimate and any helpful cautionary notes learned through pilot testing the process. This process could be applied outside of the NN/g certification context for other multiple-choice exams to meet the above-stated goals.

Phase 1: Content Analysis

Phase 1 should take place in consultation with the instructors of the associated course. In most cases, a 30-45 minute meeting can accomplish both steps while creating a shared understanding and sense of stakeholder buy-in.

Step 1: Document the basic course learning outcomes (time estimate: 15 min). This should be done in an initial meeting with all relevant course instructors who serve as subject-matter experts (SMEs). The basic learning outcomes can often be identified by looking through the existing materials and noting the various sections of the course.

Caution: The goal is to document how the course is currently being taught, not to make major changes to the course. This is not a time to discuss what changes the SMEs would like to make in the longer term, but to gain an understanding of things in their current state.

Step 2: Create a future-state test blueprint (time estimate: 15 min). This will be the blueprint for what the exam should look like given the newly defined learning objectives. It will also reflect the amount of time and attention the course should give to each of these topics. The future-state blueprint will be created as a table of specifications indicating the relationship between the course content areas identified in step one, and the relevant levels of Bloom's Taxonomy (see Appendix F).

Caution: There is a learning curve for most instructors regarding Bloom's Taxonomy. In our pilot, it tended to be more helpful to frame the Taxonomy levels with practical questions (i.e., Remember [do they know it?], Understand [do they get it?], and Apply [can they use it?]). We focused on the bottom three levels of Bloom's Taxonomy because the courses themselves were only aimed at teaching content at these basic levels.

Phase 2: Gap Identification

Phase 2 will help identify the gap between the current state of the exam and the future or ideal state of the exam described by the course instructors. This phase will prepare the way for making actual edits to the exam content.

Step 3: Assign existing exam questions to newly documented learning outcomes (time estimate: 45 min). In preparation for creating a current-state test blueprint, it is important to recognize how existing questions align with the new future-state test blueprint from step two to salvage existing work and save time.

Step 4: Determine the Bloom's Taxonomy level for each existing question (time estimate: 30 min). To determine which existing questions can be retained or adapted, it is essential to determine the level of Bloom's Taxonomy to which they align. This will reveal how some questions might be adapted to be situated at a higher or

lower level of the Taxonomy with small tweaks rather than creating entirely new questions.

Caution: Analyzing questions and determining their Bloom's Taxonomy level requires a robust familiarity with the Taxonomy to accurately categorize questions that may focus on unfamiliar content areas.

Step 5: Create a current-state test blueprint (time estimate: 2-3 hours). The purpose of the current-state test blueprint is to facilitate a comparison with the future-state test blueprint created by the instructors. The gaps between the two blueprints (e.g., a different number of questions focused on a particular learning outcome, or oriented at a particular level of Bloom's Taxonomy) will reveal areas that need more or fewer questions. Filling these gaps is essential for ensuring good content validity for the exam, and for achieving the correct difficulty level. The current-state test blueprint should utilize the learning outcomes established in step one and a smiliar table of specifications template to facilitate an "apples-to-apples" comparison (see Appendix E).

Phase 3: Exam Updates

Phase 3 outlines the process for making all changes to the exam itself.

Step 6: Update the existing exam questions (time estimate: 3-6 hrs). With the gap between the current state and future state of the exam now apparent, questions can be deleted, edited, and created to bring the exam closer to the ideal state. The following four substeps correlate to the four major problems discovered in the initial project evaluation described above. Their sequence will help guide exam improvements efforts to proceed in the most efficient manner:

A. **Content coverage.** Ensure each course content area has the correct number of questions to achieve strong content validity. Questions should address a variety of topics from within that section of the course.
B. **Difficulty level.** Ensure each level of Bloom's Taxonomy has been given the correct number of questions as indicated by the future-state test blueprint. Consider adapting some existing questions to assess mastery at a different level of Bloom's Taxonomy to avoid the need to create more new questions than necessary.
C. **Discriminatory power.** Surviving questions that have an item difficulty above 95% or have a discriminatory power lower than 0.2 are plausible candidates for further updates. Begin by checking for any distractors that never or rarely get

chosen according to the distractor discrimination indices. Adapting implausible distractors can also help improve the difficulty level of a question.

D. **Multiple-choice best practices.** Once all questions have been created or edited, check for violations of recommendations for writing quality multiple-choice questions (see Appendix G). This will help avoid any unintentional easiness or trickiness unrelated to the assessment objectives. This is the final step to avoid spending time reworking the wording of any questions that will ultimately be changed or deleted.

Phase 4: Implementation

Phase 4 describes the process for the exams team to launch the new version of the exam. This phase does not require any instructor input.

Step 7: Create and distribute the new version of the exam (time estimate: 45-60 min). To facilitate a comparison between the performance of the old and new versions of the exam, it is best to create a new exam in the exam system (onlinetesting.net) with the updates.

Caution: It is important to retain the old version of the exam with the historical item statistical data to facilitate a comparison between the performance of the new exam and the old exam.

Step 8: Monitor the new exam performance (time estimate: 3-4 months). It will likely take a few months to gather enough examinee data to make sound judgments about the performance of the new test items. Plan a time in the future to analyze the performance data and make necessary adjustments to items that do not seem to be performing well based on the above-stated criteria.

Step 9: Plan to revisit the exam in the future. A robust process for maintaining the quality of exams will periodically direct exam creators to revisit exams to check for any needed updates. NN/g exams should be revisted every one to two years.

Tracking Progress

One major component of this process is a collaboration with the SMEs responsible for creating, maintaining, and delivering the courses associated with each exam. It can be challenging to help busy SMEs understand what will be required from them as you launch a project such as this and require their time and input. I found it helpful to briefly outline the phases and steps outlined above with rough time estimates to help all

stakeholders align around the required commitment. I also found it helpful to create a progress tracker to help stakeholders stay updated regarding the progress of the project and what will come next (see Appendix I).

Indicators of Success

While NN/g is interested in better differentiating between those who truly understood the course material and those who did not, they do employ a criterion-referenced exam approach (Miller, Linn, & Gronlund, 2013), meaning that all test-takers are entitled to receive the score of 100% if they earn it. However, NN/g is also looking to create exams that are difficult enough that few (if any) questions are answered correctly by all test-takers, which would indicate that those questions do not discriminate between test-takers.

For this reason, as the performance of updated exams is monitored, it will be important to establish criteria for recognizing which items are not performing satisfactorily. Quantitatively, items that do not have a stable difficulty between 70-90% and a discriminatory power above 0.2 will be flagged for a qualitative review. On the exam level, if more than 2 test-takers out of 50 fail any given exam, this will merit a deeper review of the performance of individual items on that exam.

Conclusion

In creating this process for evaluating and improving existing exams, I discovered that components of the process are also useful for the creation of brand-new exams. Instructors responsible for creating a new course will remain responsible for the creation of the associated exam. They will be encouraged to create a basic, future-state test blueprint from the beginning to help them identify which sections of the course they should be pulling questions from and what level of Bloom's those questions should be written at. Utilizing a basic test blueprint will help maintain the overall quality of exams and prevent the natural reversion back to the problematic state described at the beginning of this report.

References

Alzu'bi, M. A. (2014). The Extend of Adaptation Bloom's Taxonomy of Cognitive Domain In English Questions Included in General Secondary Exams. Advances in Language and Literary Studies, 5(2), 67-72.

Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals. Cognitive domain.

Chase, C. I. (1999). Contemporary Assessment for Educators. Addison-Wesley Longman, 1185 Avenue of the Americas, New York, NY 10036.

Daniels, V. S. (2011). Assessing the value of certification preparation programs in higher education. American Journal of Business Education (AJBE), 4(6), 1-10.

Getto, G., & Beecher, F. (2016). Toward a model of UX education: Training UX designers within the academy. IEEE Transactions on Professional Communication, 59(2), 153-164.

Gray, C. M. (2014). Evolution of design competence in UX practice. In Proceedings of the SIGCHI Conference on human factors in computing systems (pp. 1645-1654).

Gray, C. M. (2016). "It's More of a Mindset Than a Method" UX Practitioners' Conception of Design Methods. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 4044-4055).

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

Measurement and Assessment in Teaching, M. David Miller, Robert L. Linn, Norman E. Gronlund, 11th Edition (2013), Pearson

Muhayimana, T., Kwizera, L., & Nyirahabimana, M. R. (2022). Using Bloom’s taxonomy to evaluate the cognitive levels of Primary Leaving English Exam questions in Rwandan schools. Curriculum Perspectives, 42(1), 51-63.

Nielsen, J. (2017, December 24). A 100-year view of user experience (by Jakob Nielsen). Nielsen Norman Group. Retrieved March 10, 2023, from https://www.nngroup.com/articles/100-years-ux/

Appendix A

The following table contains the following information for each of the nine sample exams used in the evaluation phase of this project:

- The number of "easy" questions that were answered correctly by 95% or more of test-takers.
- The percentage of total questions that were "easy."
- The average discriminatory power of all questions (excluding those which were answered correctly 100% of the time)
- The total number of questions

	Number of "easy" questions	**Percent of questions that are "easy"**	**Avg. DP (excl. 100%)**	**Total questions**
Exam 1	38	70%	0.05	54
Exam 2	32	67%	0.06	48
Exam 3	35	66%	0.09	53
Exam 4	18	53%	0.15	34
Exam 5	18	53%	0.18	34
Exam 6	19	44%	0.12	43
Exam 7	11	37%	0.20	30
Exam 8	15	32%	0.12	47
Exam 9	14	27%	0.12	52
Average	**22**	**50%**	**0.12**	**44**

Appendix B

The following table presents the number of items for one sample exam (Exam 1 in Appendix A) at varying levels of difficulty. The difficulty for each item was determined by the item difficulty which is the percentage of test-takers who select the correct answer option.

	95% or above	**85-94%**	**75-84%**	**65-74%**	**64% or below**
Number of Items	37	13	4	0	0

Appendix C

The following table presents the number of items for one sample exam (Exam 1 in Appendix A) at varying levels of discriminatory power. The discriminatory power for each item was determined by calculating the correlation between the number of test-takers selecting the correct answer option and their performance on the exam overall.

	0.35 or above	**0.25-0.34**	**0.15-0.24**	**0.05-0.14**	**0.04 or below**
Number of items	0	1	7	19	21

Appendix D

The following table presents the number of items for one sample exam (Exam 1 in Appendix A) at varying levels of Bloom's Taxonomy for the Cognitive Domain (Bloom, 1956; Krathwohl, 2002). I determined the Bloom's Taxonomy level based on a qualitative analysis.

	Remember	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
Number of Items	29	14	11	0	0	0

Appendix E

The following table presents the number of items for one sample exam (Exam 1 in Appendix A) which were flagged as violating at least one best practice for writing multiple-choice items (Miller, Linn, & Gronlund, 2013). I determined whether the best practices were violated based on a qualitative analysis.

	No meaningful problem	**Irrelevant material in stem**	**Long answer options**	**Variable option lengths**	**Verbal clues**	**Includes absolutes**
Number of items	3	5	8	5	6	1

Appendix F

The following template can be used to create a new table of specifications. This template is well-suited for creating both a future-state test blueprint and a current-state test blueprint. Using two versions of the same template for both of these blueprints facilitates easy comparison.

Test Blueprint Template					
Learning Outcomes	**Remember** (*Do they know it?*)	**Understand** (*Do they get it?*)	**Apply** (*Can they use it?*)	Total **number** of items	Total **percent** of items
				0	0%
				0	0%
				0	0%
				0	0%
				0	0%
				0	0%
				0	0%
				0	0%
				0	0%
				0	0%
				0	0%
				0	0%
Total **number** of items	0	0	0	**0**	0%
Total **percent** of items	0%	0%	0%	0%	

When created in a standard spreadsheet, the columns and rows on the right and bottom of the table should be filled in with formulas that will automatically total the number of questions in each content area (row) and Bloom's Taxonomy level (column). This will allow the total percentage column and row to automatically calculate the percentage of total items found in each category to give a larger overview of the structure of an exam.

Appendix G

The following table served as a shared resource for all instructors (SMEs) as they created and updated exam questions. It also served as a set of standard criteria which could be referenced as I made suggested edits to questions through the revision process.

	Applicable to all NN/g exam questions	**Yes**	**No**
1	**Does each item stem present a meaningful problem?**		
2	**Are the item stems free of irrelevant material?**		
3	**Are the alternative answers brief and free of unnecessary words?**		
4	**Are the alternatives similar in length and form?**		
5	**Are the items free of verbal clues to the answer?**		
6	**Have absolutes such as "always" or "never" been avoided, if possible?**		
7	Are the item stems stated in positive terms (if possible)?		
8	If used, has negative wording been given special emphasis (e.g., capitalized)?		
9	Are the alternatives grammatically consistent with the item stem?		
10	Is there only one correct or clearly best answer?		
11	Are the distractors plausible to low achievers?		
12	Are numerical alternatives in numerical order?		
13	Have *none of the above* and *all of the above* been avoided (or used sparingly and appropriately)?		
14	If revised, are the items still relevant to the intended learning outcomes?		
15	Have the items been set aside for a time before reviewing them?		

Reference: *Measurement and Assessment in Teaching,* M. David Miller, Robert L. Linn, Norman E. Gronlund, 11th Edition (2013), Pearson

Appendix H

The following questions can be useful for helping guide instructors (SMEs) through the process of deciding how to allocate test questions across content areas and levels of Bloom's Taxonomy in the future-state test blueprint they will create. In this project, I found that only basic versions of these questions were necessary. However, for less experienced instructors, or teams more focused on true learning objectives in addition to content areas, full versions of these questions can be helpful.

Developing Learning Objectives

1. What would you hope conference attendees **could do** by the end of the course?
    a. In what contexts should they be able to do this?
    b. To what degree of proficiency should they be able to do this?
    c. What nuances of a skill are not necessary to address in a 1-day course?

2. What would you hope conference attendees **would understand** by the end of the course?
    a. What are some central terms/rationales/principles they need to understand?
    b. To what depth do they need to know or understand these things?
    c. What common misconceptions need to be addressed?
    d. How would you know if they correctly understood these things?

3. What would you hope conference attendees **would feel** by the end of the course?
    a. What changes in priorities would you hope to see?
    b. What types of worries should be overcome?
    c. What would you want attendees to feel confident in by the end of the course?

Deciding Topic Categories

1. Which topic areas deserve the most attention to meet the learning objectives?
    a. What might be some subtopics within each larger category?
    b. Which subtopics are more important than others?

2. What topics might be "nice to know," but not essential?

3. How might you order the topic categories from most to least important?

4. How familiar do attendees tend to be with each topic?
    a. What topics are generally new or novel for many conference attendees?
    b. What topics are attendees likely already somewhat familiar with?

Distributing Questions Across Bloom's Taxonomy

1. How could attendees demonstrate their mastery of a topic?
2. How deeply and thoroughly was this topic addressed?
3. How much hands-on practice did attendees get with this topic?
4. To what degree did attendees see examples of this principle being applied/implemented?
5. To what extent was the rationale behind this topic/strategy explained and discussed?

Appendix I

This table was presented at the top of the shared documentation between the Exams Team and instructors for a course to create a shared understanding of the project's progression. This table could be accompanied or linked to a more full description of each step if helpful.

[Exam Name]	**Status**
Step 1: Document the basic course content areas (time estimate: 15 min)	Not started/ In progress/ Complete
Step 2: Create a future-state test blueprint (time estimate: 15 min)	Not started/ In progress/ Complete
Step 3: Assign existing exam questions to newly documented course content areas (time estimate: 45 min)	Not started/ In progress/ Complete
Step 4: Determine the Bloom's Taxonomy level for each existing question (time estimate: 30 min)	Not started/ In progress/ Complete
Step 5: Create a current-state test blueprint (time estimate: 2-3 hours)	Not started/ In progress/ Complete
Step 6: Update the existing exam questions (time estimate: 3-6 hrs)	Not started/ In progress/ Complete
Step 7: Create and distribute the new version of the exam (time estimate: 45-60 min)	Not started/ In progress/ Complete
Step 8: Monitor the new exam performance (time estimate: 3-4 months)	Not started/ In progress/ Complete