Technical Report

Evaluation of Question Bank-Only Generic Fundamentals Exam

Date: December 20, 2017

Prepared by: Stephanie Morrow, Ph.D.
Human Factors and Reliability Branch
Division of Risk Analysis
Office of Nuclear Regulatory Research

Prepared for: Jacob Dolecki
Operator Licensing and Training Branch
Division of Inspection and Regional Support
Office of Nuclear Reactor Regulation

Table of Contents

List of Figures

List of Tables

Executive Summary

The generic fundamentals examination (GFE) is a test administered by the Nuclear Regulatory Commission (NRC) to individuals who intend to apply for a reactor operator or senior reactor operator license at a U.S. nuclear power plant. The GFE tests an examinee's knowledge of nuclear power plant fundamentals, and must be satisfactorily completed by achieving a score of 80 percent or higher. Unless a waiver is granted, the GFE must be completed within 24 months of the date of applying for an operator license. Currently, each administration of the GFE consists of 50 multiple choice questions, of which 40 items are taken directly from a publicly available question bank maintained by the Nuclear Regulatory Commission (NRC), 5 items are developed by significantly modifying items from the question bank, and 5 items are newly developed by the contractor who administers the GFE for the NRC.

At the request of the Office of Nuclear Reactor Regulation (NRR), staff in the Office of Nuclear Regulatory Research (RES) explored whether the validity, reliability, and integrity of the GFE could be maintained if the examination was made up entirely of test items taken from the NRC's GFE question bank. In addition, if exclusive use of the GFE question bank was feasible, RES was asked to determine how many items the NRC's GFE question bank would need to contain in order to ensure examination integrity. RES staff also gathered information on computer-based testing to explore the feasibility of transitioning the GFE to a computer-based format.

The RES assessment concludes that GFEs derived solely from items in the GFE question bank would decrease the validity and reliability of the exam and cannot assure examination integrity. An exam derived entirely from disclosed question bank items would likely reduce the cognitive level of test questions to recall, increase the predictability of the exam, and decrease the overall level of difficulty of the exam. The RES analysis of GFE performance data revealed that a significantly higher proportion of individuals taking the GFE answer question bank items correctly as compared to new or modified items. Further, when new items were reused on subsequent exams they had a decrease in difficulty of nearly 12 percent. On average, 12 percent more examinees answered a question correctly when it was reused, as compared to the percentage of examinees who answered correctly the first time the question was used on an exam. These results suggest that recall and familiarity with question bank items has a significant impact on overall performance on the exam.

Based on advances in computer technology over the past two decades and prevalence of computer-based testing, linear computer-based testing appears to be a viable alternative to the paper-based GFE. Exclusive use of the GFE question bank is also not a necessary prerequisite for transitioning to computer-based testing. Other options exist for developing new exam items and ensuring item security to protect the integrity of the exam.

Regardless of how the exam is administered, accepted standards for psychological testing and research on past GFE performance provide strong evidence against exclusive use of the disclosed GFE question bank to develop new exams. Further, given that the GFE question banks are and have been disclosed to the public, the size of the question bank should not be used as a justification for deriving all examination items from the question bank. Over-reliance on the question bank would compromise the validity of the GFE as an appropriate decision-making tool for discriminating among examinees who have and have not mastered the fundamental knowledge required to operate a nuclear power reactor safely.

1. Introduction

The Generic Fundamentals Examination (GFE) is the first in a series of examinations administered to individuals who intend to apply for a reactor operator (RO) or senior reactor operator (SRO) license at a U.S. nuclear power plant under the regulatory authority of the Nuclear Regulatory Commission (NRC). Each new applicant must satisfactorily complete the GFE for the applicable reactor type (boiling- or pressurized-water reactor) within 24 months before the date of application for a license or attempt to seek a waiver. The exam covers nuclear power fundamentals identified in NRC regulations 10 CFR 55.41 and 10 CFR 55.43: components, reactor theory, heat transfer, thermodynamics, and fluid mechanics. Achieving a passing grade on the GFE provides confidence that applicants possess basic knowledge about the fundamentals of nuclear power plant operations, which is a necessary foundation for later control room related problem-solving, troubleshooting, and decision-making. The satisfactory completion of the GFE is therefore an essential component in evaluating the applicant's ability to safely and competently operate the nuclear power plant.

Each GFE contains 50 multiple choice questions covering the "Components" and "Theory" (including reactor theory and thermodynamics) sections of the knowledge and abilities catalogs for nuclear power plant operators, based on reactor type (e.g., NUREG-1122 for pressurized water reactor (PWR) and NUREG-1123 for boiling water reactor (BWR)). As currently outlined in NUREG-1021, "Operator Licensing Examination Standards for Power Reactors," the examination includes 40 questions taken directly from the NRC's GFE question bank for the applicable vendor type (bank items), 5 questions that are derived from existing bank questions by making one or more significant modifications (modified items), and 5 questions that are newly developed (new items). According to NUREG-1021, in order for a question to be considered significantly modified, a change must be made to at least one pertinent condition in the stem and at least one distractor. The question bank is publicly available at the NRC's website (https://www.nrc.gov/reactors/operator-licensing/generic-fundamentals-examinations.html). Questions are added to the website on a delayed schedule, such that the publicly-available bank excludes the two most recent examinations.

One proposal under consideration is to modify the current structure of the GFE to have all 50 GFE questions come directly from the NRC's GFE question bank. The current size of the GFE question bank (i.e., approximately 2000 questions in each of the PWR and BWR GFE banks) has motivated this proposed change. Thus, a fundamental question to be addressed before this potential change can be considered is whether the existing GFE question bank is sufficient, such that additional new questions are not necessary to ensure a discriminating, valid, and reliable examination.

The Office of Nuclear Reactor Regulation (NRR) requested technical assistance from the Office of Nuclear Regulatory Research (RES) to evaluate this proposed change to the structure of the GFE. In particular, RES staff was asked to determine if an examination with all of the test items taken directly from the NRC's GFE question bank can maintain test validity, integrity, and reliability, and, if so, how many items the NRC's GFE question bank needs to contain in order to ensure examination integrity. This report describes the assessment performed by RES staff regarding the proposed change to a question bank-only GFE.

2. Technical Basis for Testing

Standardized psychological testing in the workplace is generally used as a means of measuring knowledge, skills, and abilities thought to be important in eventual performance of a job. The basic approach to group-administered multiple-choice testing was first developed by the U.S. military during World War I (National Research Council, 1991). The early military studies focused on tests that succeeded in reducing the number of dropouts from pilot training, and over time this approach yielded considerable success – 77 percent of the applicants with the highest scores succeeded in the training course. Subsequent development of screening and classification measures has been extended throughout public and private industry, and the resulting knowledge base has helped to shape psychological testing research more broadly. Psychological testing is used today in a variety of job-related applications, such as employee selection, classification, certification, and licensing.

2.1. Standards for Psychological Testing

The Standards for Educational and Psychological Testing is a guidance document that outlines explicit criteria for the development and evaluation of tests and testing practices. The Standards are jointly produced by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA/APA/NCME, 2014) to promote the sound and ethical use of tests.

The foundation for justifying use of a psychological test is grounded in establishing evidence of validity and reliability. The Standards define validity as the degree to which evidence and theory support the interpretations of test scores for the proposed use of the test. Note that validity is not a property of the test itself; validity is based on how the test is intended to be used and interpreted. The process of validation involved accumulating relevant evidence to provide a sound scientific basis for how the test will be interpreted. Reliability is a prerequisite for validity and refers to the consistency of test scores across replications of a test. The standardization of the testing process helps create consistency of measurement. The need for precision in testing increases as the consequences of decisions and interpretations grow in importance.

Adequate evidence of validity often requires multiple sources based on how the test will be used. The Standards discuss various sources of evidence that can be used to evaluate the validity of a proposed interpretation of test scores for a particular use:

- Evidence based on *test content* examines the relationship between the content of the test and the construct it is intended to measure.
- Evidence based on *response process* examines the fit between the construct being measured by the test and the actual cognitive processes engaged in by test takers.
- Evidence based on the *internal structure* of a test examines the relationships among test items, such as item difficulty.
- Evidence based on *test-criterion relationships* examines how well the test scores predict performance.

The Standards also discuss the importance of considering potential unintended consequences of a test, including construct-irrelevant variance or construct contamination (the degree to which tests scores are affected by processes that are irrelevant to the test's intended purpose). Test scores may be systematically influenced to some extent by response processes that are not part of the construct. For example, quality of handwriting may influence how a written essay is

graded, but it is irrelevant to judging the content of the essay. Differences in reading skill may influence performance on a written test of mechanical comprehension, but may be unrelated to an individual's mechanical comprehension abilities. Similarly, prior knowledge of test questions may change an individual's cognitive processing from analysis to simple recall to arrive at the correct answer. As a result, examinees' test scores may not reflect actual mastery of the content domains, but rather memorization and recall ability.

2.2. Regulatory Basis

In the case of operator licensing, the GFE is a key measure that allows the NRC to make a confident decision regarding the safety-significant performance of the individual seeking a license (NUREG-1021, Appendix B). Both the internal and external attributes of the GFE can impact validity and reliability. Internal attributes include such things as the source of the content for the questions, level of knowledge, level of difficulty, and extent of the use of question banks. External attributes include the length of the examination, security procedures, proctoring instructions, and other administrative details. If the internal and external attributes of examinations are allowed to vary significantly, the *uniform conditions* that are required by Section 107 of the *Atomic Energy Act of 1954*, as amended, and the basis upon which the NRC's licensing decisions rest are challenged. Further, the NRC must reasonably control and structure the examination processes to ensure the integrity of the licenses it issues (NUREG-1021, Appendix A).

2.3. Consequences for Operator Performance

The GFE is an example of a high-stakes test, where the decisions made as a result of the test have important consequences for both the examinee (i.e., ability to get a license to obtain a job as a reactor operator) and the public (i.e., assurance that nuclear power plants are operated safely by competent personnel). NUREG-1021, Appendix B discusses the potential implications of underestimating the importance of knowledge testing for subsequent operator performance:

> *"Deemphasizing or sidestepping knowledge testing through careless or simplistic testing processes or treating it secondarily to other portions of the examination that are more operationally oriented could affect subsequent job performance. Failing to focus on testing the individual operator's cognitive abilities (i.e., comprehension, problem-solving, and decision-making) or paying insufficient attention to the operator's fundamental understanding of job content (e.g., systems, components, and procedures) may ultimately place job performance at risk of gradual degradation. When the demand for disciplined learning and study declines or the level of knowledge (depth of application) required for the job is reduced, it could lead to less time spent in training preparation, less mental review and practice, more forgetting of factual details, less reinforcement and application of job concepts, and a gradual decline in performance.*
>
> *Moreover, without a solid fundamental knowledge base, operators may not perform acceptably in situations that are not specifically addressed in procedures. Since every performance has an underlying knowledge component, that knowledge and its depth need development and assessment to ensure the operators' competence on the job. Studies assessing mental performance in cognitively demanding emergencies point out that higher level cognitive thought (such as event diagnosis and response planning) are important in responding to safety-related events [NUREG/CR-6208]."*

3. Impact of Question Bank Use on Evidence of GFE Validity

The purpose of the GFE is to establish that examinees have sufficient mastery of the basic fundamentals of nuclear power plant operations. Evidence of the validity of the GFE for making licensing decisions is based on the test content, the cognitive processes used to answer the test questions, internal structure, and criterion used to determine acceptable performance. This section discusses the various source of evidence and how exclusive use of a question bank would affect the validity of the GFE.

3.1. GFE Content and the Validity Inference

The primary goal when using psychological tests for job qualification is to devise a test that can predict on-the-job performance by using test items that are representative of the knowledge, skills, and abilities needed to perform the job. This is commonly referred to more simply as knowledge and abilities (K/As) in the Operator Licensing Program. Because a single test cannot measure all of the knowledge required to be a licensed operator, the test questions must be constructed based on a sample of the required knowledge in a manner that allows inferences to be made regarding the examinees' performance on the broader population of knowledge, even though it was not tested. This is referred to as a "validity inference" (NUREG-1021, Appendix A). The sample must be carefully selected to be representative of the relevant content domains in order to conclude that the untested knowledge is proportionately known or not known in relation to the score on the sample. When a sample is biased or skewed in a particular direction, it introduces sampling error, which makes it inappropriate to infer or generalize that the examinees have mastered the larger population of untested knowledge from which the sample was drawn (NUREG-1021, Appendix A).

The standardized sampling plan used to develop the GFE ensures that each exam tests a representative sample of fundamental knowledge for nuclear reactor operations. Figure 1 presents a conceptual model of how test questions are sampled to be representative of the GFE content domains. The current GFE is composed of 50 questions sampled from three primary content domains: components, reactor theory, and thermodynamics. Each content domain is divided into multiple topic areas based on the knowledge and ability (K/A) catalogs for Pressurized Water Reactors (PWRs), and Boiling Water Reactors (BWRs), Westinghouse AP1000 Pressurized Water Reactor, and Advanced Boiling Water Reactor (i.e., NUREG-1122, NUREG-1123, NUREG-2103, or NUREG-2104). NUREG-1021, Examination Section (ES)-205, specifies how many questions on the GFE must come from each topic area.

Use of new and modified test items ensures that the entire sample is not known prior to the exam and minimizes the overall predictability of the exam. One implication of using a static question bank, where possible test items do not change and are known in advance to the examinees, is a violation of the validity inference. Rather than sampling from the content domains, the test items become the population of knowledge that examinees must recall to pass the exam. Content that is not covered by an existing test item is no longer necessary to study. Effectively, the knowledge that examinees must have is reduced to the specific test questions (i.e., the white circles in Figure 1), rather than whole of the content domains (i.e., the blue area in Figure 1). The unpredictability of new and modified test questions provides a level of protection against passing examinees who only study questions from previous exams at the expense of studying the content domains from which the exam is intended to sample.

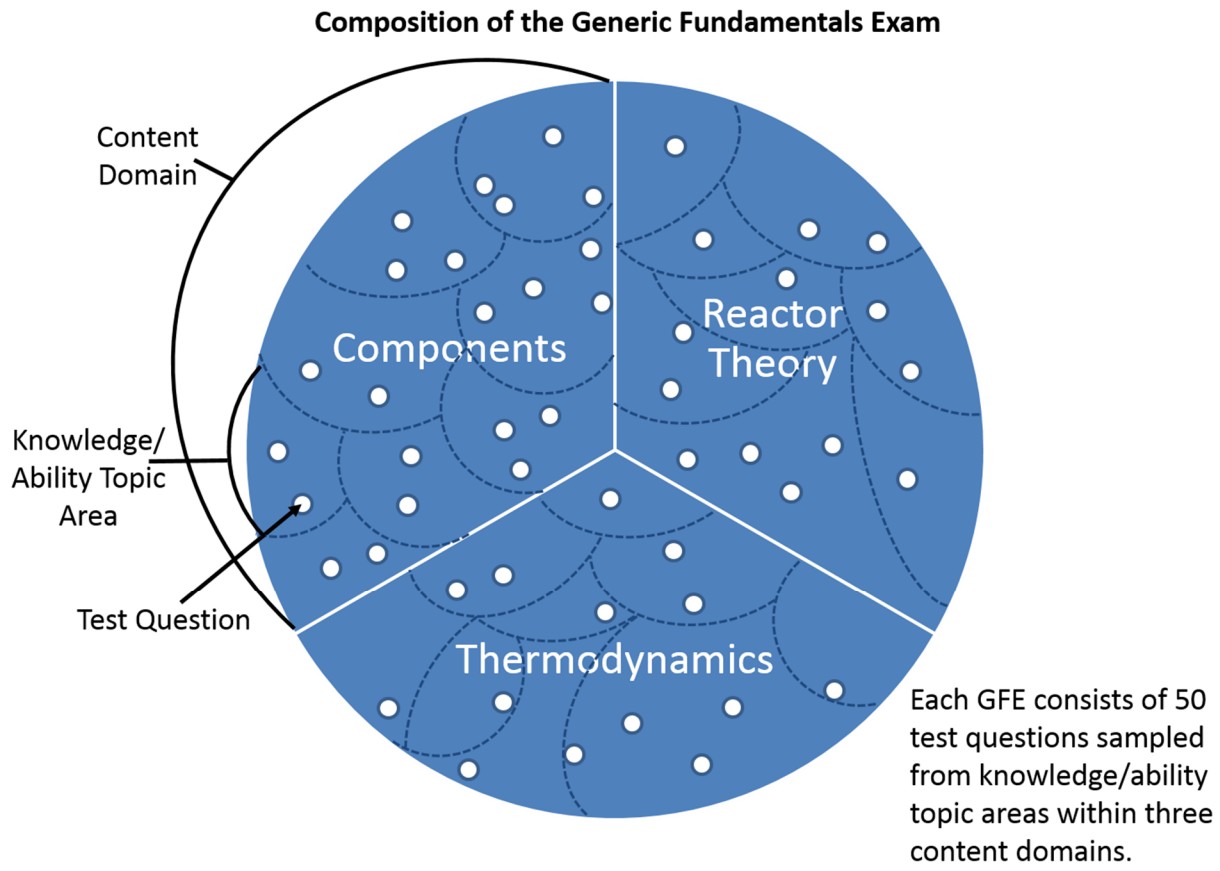


Figure 1. Conceptual Model of the Composition of the Generic Fundamentals Exam

NUREG-1021, Appendix A also references the potential implications of copying all or a significant portion of test items from a question bank:

> *Although studying past examinations can have a positive learning value, total predictability of examination coverage through overreliance upon examination banks reduces examination integrity. When the examinees know the precise and limited pool from which test items will be drawn, they will tend only to study from that pool (i.e., studying to the test) and may likely exclude from study the larger domain of job knowledge. When this occurs, it decreases the confidence in the validity inferences that are made from performance on the test to that of the larger realm of knowledge or skill to be mastered.*

3.2. Level of Knowledge and Cognitive Response Processes

In addition to differences in content, test items can vary with regard to their level of knowledge. Level of knowledge represents the range of mental demands and cognitive response processes required to answer a question or perform a task. A test item's level of knowledge can range from retrieving fundamental knowledge (low-level) to retrieving that knowledge and also understanding, analyzing, and synthesizing that knowledge with other knowledge (high-level).

Bloom's Taxonomy is a theoretical knowledge classification system used to determine the level of knowledge that an exam question is testing (Bloom, 1956; Anderson et al., 2001). Bloom's

Taxonomy provides a framework for grouping exam questions based on the level (depth) of mental thought and performance required to answer the question. When evaluating level of knowledge, two key elements must be considered: (1) the number and type of mental steps necessary to process the given data and arrive at the correct answer, and (2) the training and experience level of the target test group (NUREG-1021, Appendix B).

NUREG-1021 uses a modified version of Bloom's Taxonomy to classify test items into one of three levels of knowledge (see Figure 2):

- Level 1 (i.e., fundamental knowledge or simple memory) tests the recall or recognition of discrete bits of information. Examples include knowledge of terminology, definitions, set points, patterns, structures, procedural steps and cautions, and other specific facts.
- Level 2 (i.e., comprehension) involves the mental process of understanding the material by relating it to its own parts or to some other material. Examples include rephrasing information in different words, describing or recognizing relationships, showing similarities and differences among parts or wholes, and recognizing how systems interact, including consequences or implications.
- Level 3 (i.e., analysis, synthesis, or application) testing is a more active and product-oriented testing approach, which involves the multifaceted mental process of assembling, sorting, or integrating the parts (information bits and their relationships) to predict an event or outcome, solve a problem, or create something new. This level requires mentally using the knowledge and its meaning to solve problems.

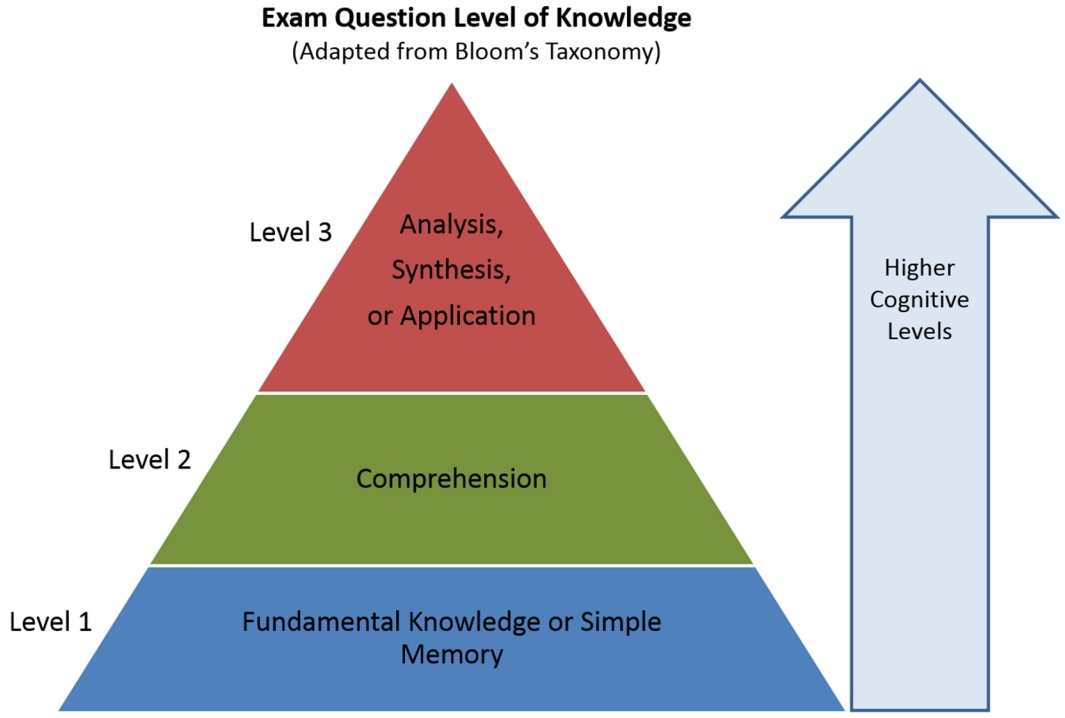


Figure 2. Levels of Knowledge for Exam Questions (Adapted from Bloom's Taxonomy)

Test items that are more relevant to actual job performance demonstrate greater evidence of validity. Test items can be relevant to job performance based on their content (e.g., questions about reactor theory), but also based on their level of knowledge. For instance, there are a number of high-level cognitive activities central to the job performance of reactor operators, including analysis, prediction of events or system responses, and problem-solving. Test items that are written at a high cognitive level are generally more relevant to job performance than test items written at a low cognitive level because they require examinees to perform cognitive activities that are similar to what they would perform on the job (i.e., comprehension, analysis, synthesis, and application of knowledge).

When questions are known to examinees in advance it increases probability that examinees will select the correct answer because they recognize the question from their studies, rather than because they comprehend the material. This would effectively reduce the level of knowledge of the test questions because examinees only use the cognitive response processes of recognition and recall to select the correct answer. NUREG-1021, Appendix A discusses the possible reduction in level of knowledge as a result of over-reliance on question banks:

> *Previously administered test items reduce examination integrity because examination discrimination is reduced. Discrimination is reduced because the cognitive level at which the examinees are tested could decrease to the simple recognition level if the item bank is small and available for the examinees to study. The comprehension and analysis levels of knowledge may not be assessable because mental thought has been reduced to a recognition level, and decision-making is absent because test items have been rehearsed and are anticipated. In short, challenge and mental analysis are lost and the examinees are tested at a role-rehearsal level. An examination cannot assess higher cognitive and analytical abilities if a significant portion of the items within the examination have already been seen.*

3.3. Level of Difficulty and Internal Structure

Test items can also vary in terms of level of difficulty. An item's level of difficulty is separate from its level of knowledge. For example, an item may test at a low cognitive level but high difficulty level by requiring an examinee to recall a specific, but obscure fact about a power reactor. Item difficulty can be assessed based on the percentage of examinees that are expected to answer the question correctly. For the cut score of 80 percent to be meaningful, each exam should include a balance of more and less difficult items near that difficulty level. Level of difficulty is one of the factors that GFE developers consider when evaluating whether a question is acceptable for inclusion in the exam. NUREG-1021, Appendix B recommends a target level of difficulty in the range of 70 to 90 percent for individual test items (i.e., 70 to 90 percent of examinees are expected to answer the question correctly).

An exam can include an appropriate distribution of questions from the relevant content domains, but if the majority of questions are too difficult or too easy, then the exam may not appropriately discriminate among examinees. Test items that are so difficult that few examinees are expected to answer correctly do not discriminate. Likewise, test items that are so easy or fundamental that even those examinees that are known to have performance problems will be able to answer correctly do not effectively discriminate.

One factor that may affect the difficulty level of an item is an examinee's familiarity with the same or similar questions. Questions that are taken from a disclosed question bank are more

likely to be familiar to examinees because they encountered the questions while practicing and preparing for the exam. The use of only question bank items to construct an exam may have the effect of decreasing the overall difficulty of the exam and thereby decrease the exam's ability to discriminate among examinees.

3.4. Criterion-Based Interpretation of Exam Results

The GFE uses a minimum cut score or grade of 80 percent as the criterion that an examinee must achieve to pass the exam. The underlying expectation in a criterion-referenced test is that if examinees can correctly answer a particular minimum number of questions from the test (based on the established cut score), then they can reasonably be expected to possess an adequate level of the knowledge and abilities from which the test sample was based. In other words, passing the GFE indicates that the examinee has demonstrate sufficient knowledge of the fundamentals of safe operation of a nuclear power plant and may proceed in the site-specific licensing process. For a criterion-referenced test to be effective, both the individual test items and the overall examination must discriminate between examinees who have and have not mastered the required knowledge, skills, and abilities.

In the writing of test items, the exam developers are also well aware of the NRC-established passing score of 80 percent. They use this knowledge to control the nature and difficulty of the examination, such that an examinee who is deemed to be qualified scores above the passing grade, while an examinee who is deemed to be unqualified scores below that grade. Rather than explicitly judging the probability that a minimally qualified examinee will pass an item, the developers are implicitly being asked to construct an examination where, in the developer's judgment, the minimally qualified examinee will obtain a score of at least 80 percent (NUREG-1021, Appendix A).

The use of only question bank items to construct new exams may have an impact on the validity of the interpretation of the exam results. Examinees who take a test composed of only question bank items could achieve a passing score because they 1) possess an adequate level of knowledge or 2) recall a sufficient number of specific test questions. As a result, the 80 percent criterion may no longer be a meaningful criterion for decision-making if examinees can be expected to achieve that score based on recall of question bank items.

4. Use of Question Banks

Question banks can serve as an important resource for test development and to facilitate learning. Reusing items from a question bank saves substantial resources that would be needed to develop new test items for each administration of the exam. Overlap among items across exams also assists with maintaining reliability in terms of the consistency of the test items across different examinations. Use of question banks for training can help examinees become more familiar with the types of questions they can expect to encounter on the exam, and provides valuable material for examinees to use in practice to self-test their mastery of the content domains. However, over-reliance on question banks, particularly when the items in the bank are known to examinees prior to the examination, can have a detrimental effect on the reliability and validity of the examination.

4.1. Item Exposure and Item Disclosure in Question Banks

Item exposure and item security are important considerations in standardized testing (Fitzgerald & Mulkey, 2013). As a result, most standardized testing programs use question banks that are secured (i.e., not disclosed to test takers) and constantly evolving. New items are added to the question bank, pilot tested to ensure reliability and validity, and then added to the rotation of possible items for use in the testing program. The frequency with which an individual item is used in a test is controlled to avoid over-exposing an item to examinees and thereby increasing the probability that they will know the correct answer to the item based on practice rather than knowledge. In addition, items may be cycled in and out of the “active” question bank or retired from the question bank if their exposure rate is too high or the item is no longer applicable. Items are typically only released to the public once they have been taken out of use in the testing program.

The disclosure of past items and subsequent reuse of those items on future exams may compromise the validity of the exam. Advance knowledge of the test content can affect examinees’ performance, such that examinees who do not have full comprehension of the material are more likely to benefit from an exam where all possible items have been disclosed prior to the exam. As such, the examinees’ performance on the test may not necessarily be commensurate with their level of understanding of the material being tested.

4.2. Guidance on Question Bank Size

A common concern when using question banks is whether there are enough items in the question bank to support development of multiple parallel forms of the test while maintaining test security. One well-established computer-based testing organization suggests that for randomly generated on-demand testing (also called linear on-the-fly testing or LOFT), the item bank should consist of at least 10 times the number of items needed for any one form of a test (Prometric, 2017; Fitzgerald & Mulkey, 2013). Another research paper suggests multiplying the number of items used in the test by the number of anticipated test-takers to determine the size of the question bank (Burghof, 2001). However, in each case the proposed values assume that items are secured (i.e., not disclosed to test takers), and that items may be cycled in and out of the question bank to limit overall exposure of the item (i.e., the number of times the item has been used on a test).

Generally, guidance on test development does not prescribe a question bank of a particular size. Instead, the appropriate size of a question bank should be determined by considerations

such as item security and disclosure, testing frequency and population, and the significance of the test. The validity and reliability of a test can be compromised even with very large question banks if the items are reused too frequently or available to examinees prior to the test.

4.3. Previous Research on Item Disclosure and Question Bank Size

A 1980 study by Hale and colleagues directly examined the effects of item disclosure and question bank size on student performance using the test of English as a foreign language (TOEFL). They found a statistically significant effect of item disclosure on performance, such that students who had access to test items performed an average of 4.6 percentage points better on a test that included disclosed items as compared to a test that included all new items. The effect was more pronounced for students who studied from a smaller question bank. The students who had access to a question bank of 900 items had an average increase of 6.3 percentage points, whereas students who had access to a larger question bank (1800 items) had an average increase of 2.9 percentage points.

Hale et al. discuss two types of effects that can result from test preparation activities with disclosed test items: a general learning effect (i.e., improvement in performance from experience with the test in general) and a specific recall effect (i.e., improvement in performance due to recall of specific questions and answers on the test). A general learning effect should equally impact performance on tests with disclosed items and tests with only new items. Given the observed differences in scores between the tests with new and disclosed items, the study suggests that performance improvements are the result of specific recall of test items, rather than general learning. Although the recall effect seemed to decrease with a larger question bank, there was still a statistically significant increase in performance when the test included disclosed test items. In the Hale et al. study, the disclosed test materials were made available to students for an average of four weeks. It is reasonable to expect that if students were given more time with the test materials, then they would be able to study a larger item bank more carefully and thereby increase their ability to recall more test items. In addition, motivation to perform well on a test is likely to affect time spent studying. This can be a particular concern in the case of high-stakes testing for licensure or credentialing, where test performance impacts one's ability to get or keep a job. Although disclosure effects may decrease as the size of the question bank increases, the time available and motivation to study the disclosed test material is likely to increase the effects of item disclosure on performance.

4.4. Benchmark with the Federal Aviation Administration

The Federal Aviation Administration (FAA) airman knowledge test is an example of another high-stakes test in a safety-critical industry, and can be a useful benchmark against the testing practices used in the GFE for operator licensing. A frequently asked questions document published by the FAA directly addresses their current approach to the airman knowledge testing program (FAA, 2017).

> *"The FAA makes every effort to maintain the integrity and security of actual knowledge test questions through regular review and revision of the test question item bank. We have recently intensified this review and revision process, so it is increasingly unlikely that applicants will see an exact match between sample questions and actual test questions. The FAA does not publish actual knowledge test questions, in part because at least two independent studies indicate that publication of active questions could negatively affect learning and understanding, as well as undermine the validity of the*

knowledge test as an assessment tool. The agency does provide sample knowledge tests on the FAA website. The questions in these sample tests are intended to help applicants understand the scope and type of knowledge that will be tested to qualify for the target certificate or rating. The goal is for applicants to devote their efforts to mastering the fundamental aeronautical knowledge necessary for safe operations in the National Airspace System (NAS) rather than to memorizing specific questions and answers."

Prior to the FAA's current practice of securing their question banks from disclosure, the question banks were available to the public. For example, the private pilot airplane knowledge test consisted of 60 questions drawn from a static bank of 915 questions. However, over time the FAA became increasingly concerned about the potential for applicants to use rote memorization strategies to answer test questions. The FAA began using computer-based testing for its airman knowledge test in the 1990's. One advantage of the computerized administration of the test is that timing information can be automatically recorded for each test-taker. In 2002, the FAA examined the amount of time test-takers spent on each question and found evidence that many test-takers were answering questions in far less time than would be required for the average human to even read the question and possible answers (Casner et al., 2004). Questions that should have required several minutes to work through complex calculations were being answered in a few seconds, clearly suggesting that rote memorization was at work.

The FAA then sponsored a study to examine whether pilot applicants were satisfactorily learning the material being tested, or whether memorization of the question bank was at the expense of understanding the material. The researchers developed an experimental knowledge test with questions that systematically varied from the questions in the FAA question bank and administered it to private pilot students. Their goal was to discover if test-takers performed similarly on questions that were not disclosed prior to the test as compared to questions from the disclosed FAA question bank. They found that test-takers performed significantly worse on skill-based questions where different data had been substituted for the data in the original question from the FAA question bank (mean score of 87.9 percent and 73.8 percent, respectively). Test-takers also performed significantly worse on knowledge questions that they did not have an opportunity to see in advance (mean score of 87.5 percent and 64.6 percent, respectively). These results suggest that releasing test questions in advance may negatively affect learning and understanding of the knowledge and skills the test is intending to measure, and thereby reduce the validity of the test as an assessment tool. As a result of these studies, the FAA made a number of changes to their knowledge tests in 2003, including randomly shuffling answer choices, restricting the number of questions made available to the public, and using different data for skills questions that are released to the public (Casner et al., 2004).

4.5. Composition of Current GFE Question Banks

The PWR and BWR question banks each contain approximately 2,000 items. Given that each exam consists of 50 items, a 2,000 item question bank seems like a large number of items to choose from. However, each individual item is not independent from all other items. The question banks were developed over time not only by adding new questions for each test administration but also by modifying subsets of existing bank questions. Early versions of the GFE used a combination of new questions developed by subject matter experts and previously developed questions from the BWR and PWR Generic Fundamentals Question Banks published by the Institute of Nuclear Power Operations (INPO). There is also some overlap

between questions in the PWR and BWR question banks when the question relates to theory, components, or systems that are the same regardless of the specific technology used.

Exams developed in the early 1990s used higher proportions of new questions, which helped to build the available question bank. By 1994, the composition of the GFE was standardized to consist of 50 items from the established question bank, 40 modified items, and 10 new items. In 2001, the composition of the GFE changed to 80 items from the question bank, 10 modified items, and 10 new items. In 2004, the GFE was reduced to 50 test questions, with 40 items directly from the question bank, 5 modified, and 5 new items. As a result of this process of adding both new and modified items to the question banks, approximately 60 percent of the items added to the PWR and BWR question banks since 1994 are modified versions of other items in the banks. The overlap among items means that there are restrictions on randomly selecting items for inclusion in any one 50-item version of the GFE.

Further, there are specific content domain requirements for each GFE, which also limits what items can be included on any one exam. Table 1 lists the number of items that must be included in each GFE per topic area as compared to the number of items currently available in the GFE question banks for each topic area (NUREG-1021, Rev. 10).

Table 1. Required Number of GFE Items and Number of Items in the GFE Question Bank by Topic (From NUREG-1021, Rev. 10)

	PWR		BWR	
Topic	**Items Required**	**Items in Bank**	**Items Required**	**Items in Bank**
Components				
Valves	2	95	3	101
Sensors and Detectors	4	211	4	205
Controllers and Positioners	3	103	2	95
Pumps	4	173	4	181
Motors and Generators	2	103	2	111
Heat Exchangers and Condensers	2	75	3	100
Demineralizers and Ion Exchangers	2	54	2	49
Breakers, Relays, and Disconnects	3	114	2	100
Reactor Theory				
Neutrons	1	21	1	39
Neutron Life Cycle	1	57	1	42
Reactor Kinetics and Neutron Sources	1	65	1	84
Reactivity Coefficients	2	79	2	52
Control Rods	2	55	2	55
Fission Product Poisons	2	102	2	107
Fuel Depletion and Burnable Poisons	1	17	1	12
Reactor Operational Physics	4	173	4	166
Thermodynamics				
Thermodynamic Units and Properties	1	26	1	15

Steam	2	82	1	90
Thermodynamic Processes	1	42	1	33
Thermodynamic Cycles	1	23	1	20
Fluid Statics and Dynamics	2	93	2	78
Heat Transfer	1	39	1	59
Thermal Hydraulics	4	120	3	112
Core Thermal Limits	1	35	3	95
Brittle Fracture and Vessel Thermal Stress	1	70	1	45
TOTAL ITEMS	**50**	**2027**	**50**	**2046**

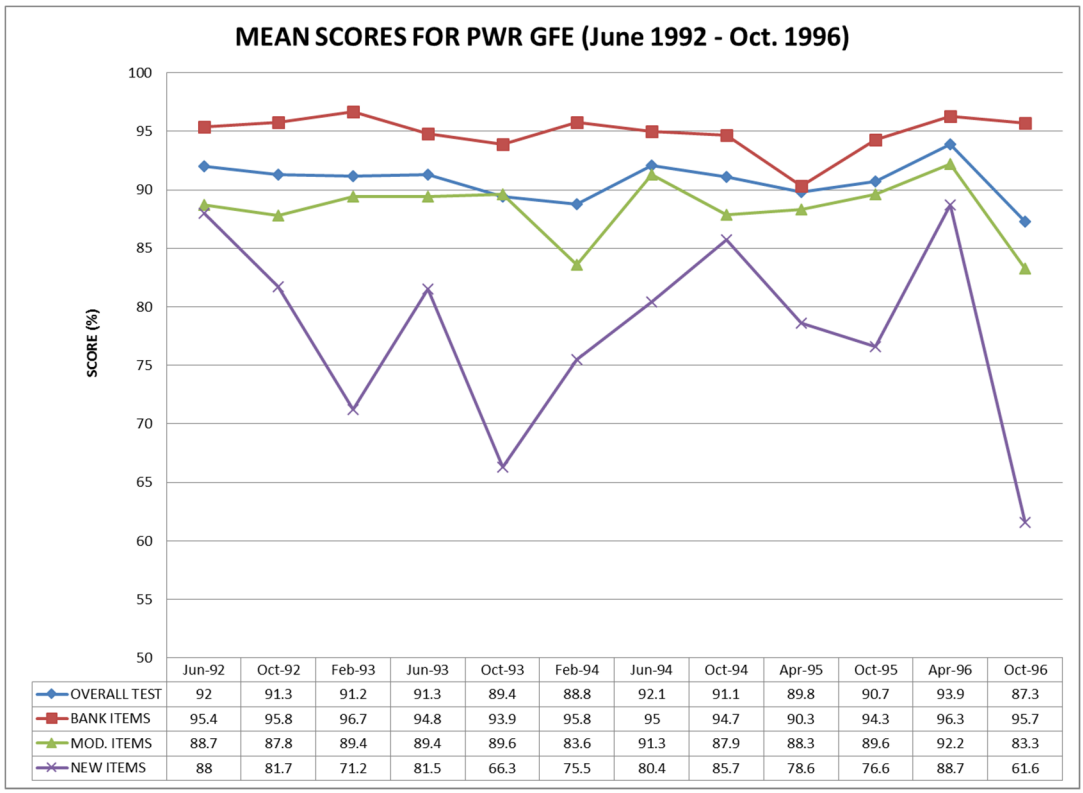
Even in cases where a topic area has a large number of possible items (e.g., Reactor Operational Physics has 173 possible items in the PWR bank, and 166 possible items in the BWR bank), many of the items were created by modifying existing bank items and therefore cannot be used within the same exam. Topic areas with few items in the question bank (e.g., Fuel Depletion and Burnable Poisons has 17 possible items in the PWR bank, and 12 possible items in the BWR bank) are even more susceptible to recall because examinees can expect that at least one of those questions will appear on the exam.

In addition, there are cases where questions from different topic areas are related, such that a question from one topic area provides cues that would help examinees answer questions from other topics areas. Conversely, interdependent questions may create “double jeopardy” situations where examinees who get one question wrong are very likely to get the other question wrong.

Any of these situations would decrease the validity of the exam as a fair assessment of an examinee’s knowledge, skills, and abilities. As a result, one cannot develop a valid 50 item test using a simple random sample from the 2,000 items in the respective PWR and BWR question banks. Instead, the sampling must be stratified based on content domain, interdependence among questions, and level of difficulty, which reduces the overall number of items available to create multiple unique exams from the question bank.

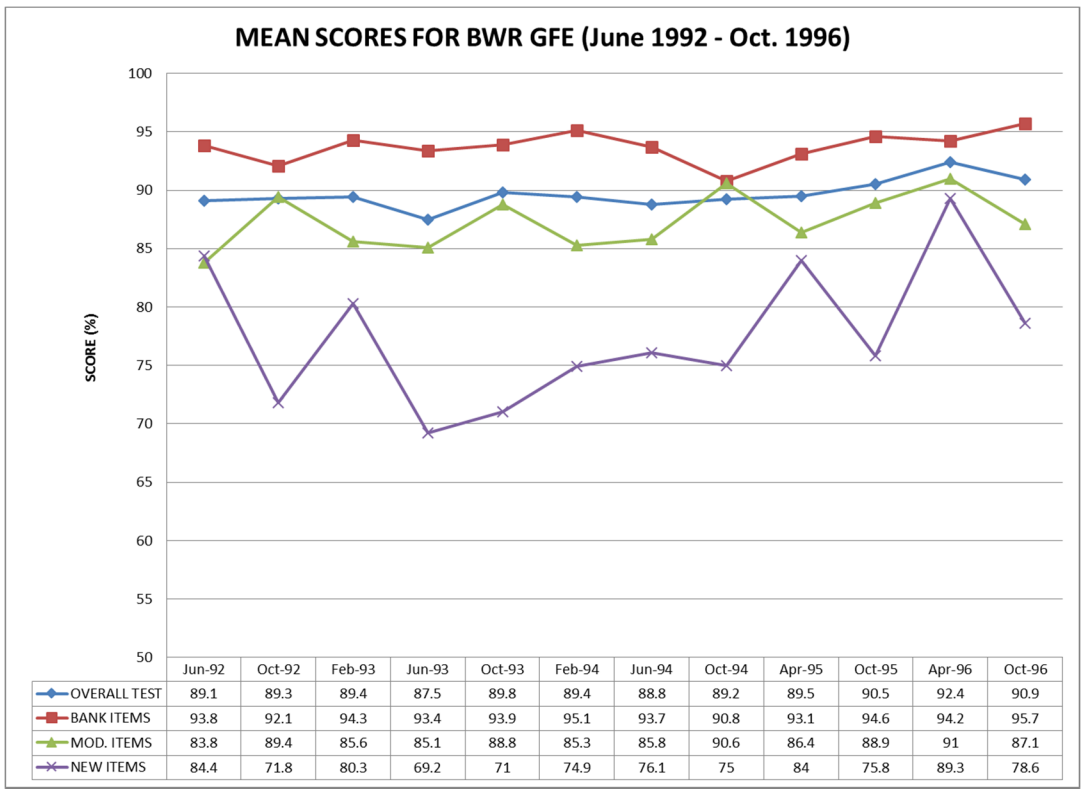
4.6. Research Comparing New, Modified, and Bank Items in the GFE

One of the primary concerns with using disclosed question bank items on new exams is that question bank items are less able to discriminate among examinees who have and have not mastered the content domains being tested. A 1997 study by former NRC staff, Dr. George Usova, examined scores on the GFE based on item type (i.e., reused question bank items, modified question bank items, and new test items) to determine whether there were differences in the discrimination ability of different item types. He analyzed data collected from 1991-1996 that included 28 examinations and 2064 examinees. At that time, the GFE question banks consisted of 1136 items for the PWR exam and 1155 items for the BWR exam. The data reported by Dr. Usova are reproduced in Figure 3 and Figure 4, and show a clear pattern of results where reused question bank items consistently demonstrated the highest mean scores, modified question bank items had slightly lower mean scores, and new test items consistently had the lowest mean scores. In other words, examinees were consistently more likely to answer reused test items correctly and less likely to answer new test items correctly.



	Jun-92	Oct-92	Feb-93	Jun-93	Oct-93	Feb-94	Jun-94	Oct-94	Apr-95	Oct-95	Apr-96	Oct-96
OVERALL TEST	92	91.3	91.2	91.3	89.4	88.8	92.1	91.1	89.8	90.7	93.9	87.3
BANK ITEMS	95.4	95.8	96.7	94.8	93.9	95.8	95	94.7	90.3	94.3	96.3	95.7
MOD. ITEMS	88.7	87.8	89.4	89.4	89.6	83.6	91.3	87.9	88.3	89.6	92.2	83.3
NEW ITEMS	88	81.7	71.2	81.5	66.3	75.5	80.4	85.7	78.6	76.6	88.7	61.6

Figure 3. Mean Scores on the PWR GFE (June 1992 – October 1996) from Usova, 1997.

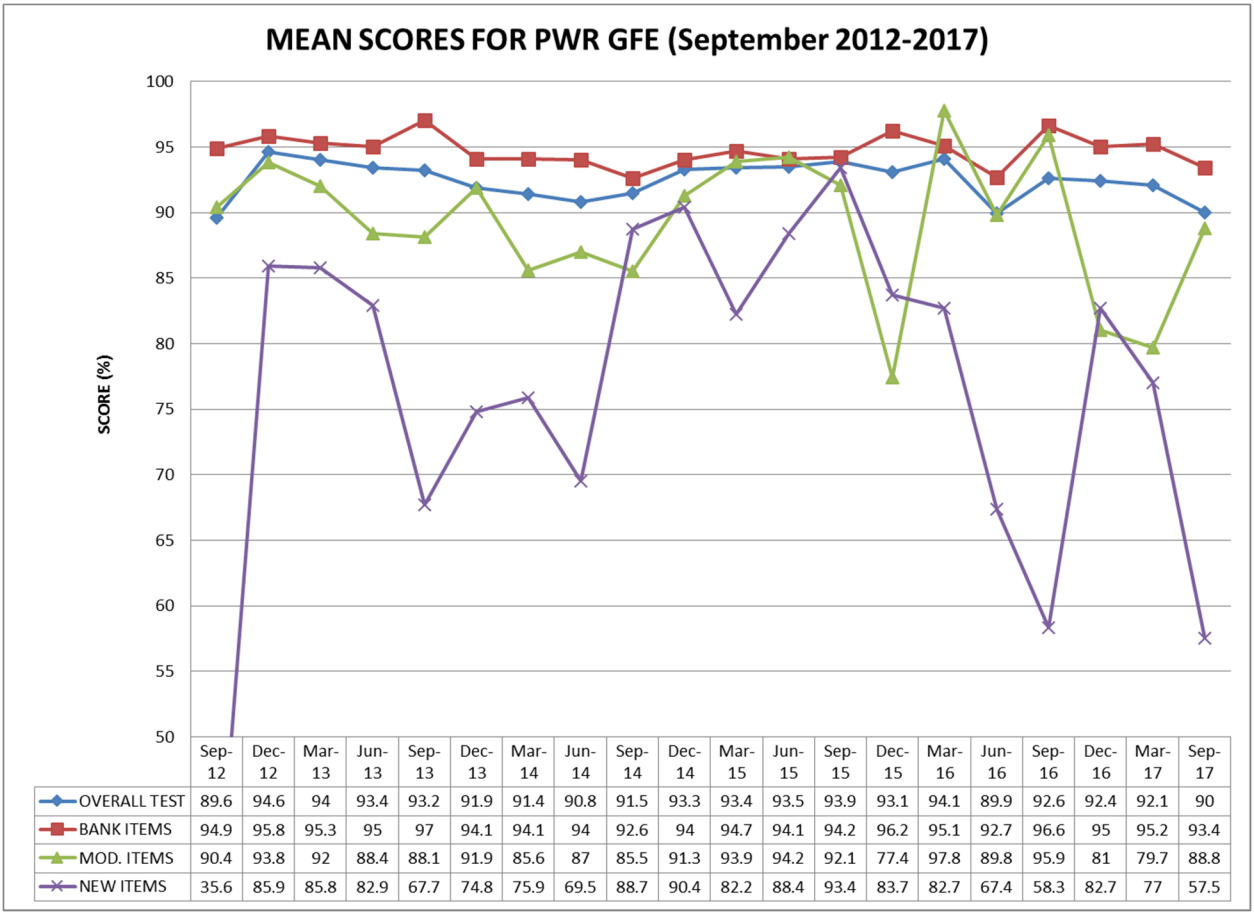


	Jun-92	Oct-92	Feb-93	Jun-93	Oct-93	Feb-94	Jun-94	Oct-94	Apr-95	Oct-95	Apr-96	Oct-96
OVERALL TEST	89.1	89.3	89.4	87.5	89.8	89.4	88.8	89.2	89.5	90.5	92.4	90.9
BANK ITEMS	93.8	92.1	94.3	93.4	93.9	95.1	93.7	90.8	93.1	94.6	94.2	95.7
MOD. ITEMS	83.8	89.4	85.6	85.1	88.8	85.3	85.8	90.6	86.4	88.9	91	87.1
NEW ITEMS	84.4	71.8	80.3	69.2	71	74.9	76.1	75	84	75.8	89.3	78.6

Figure 4. Mean Scores on the BWR GFE (June 1992 – October 1996) from Usova, 1997.

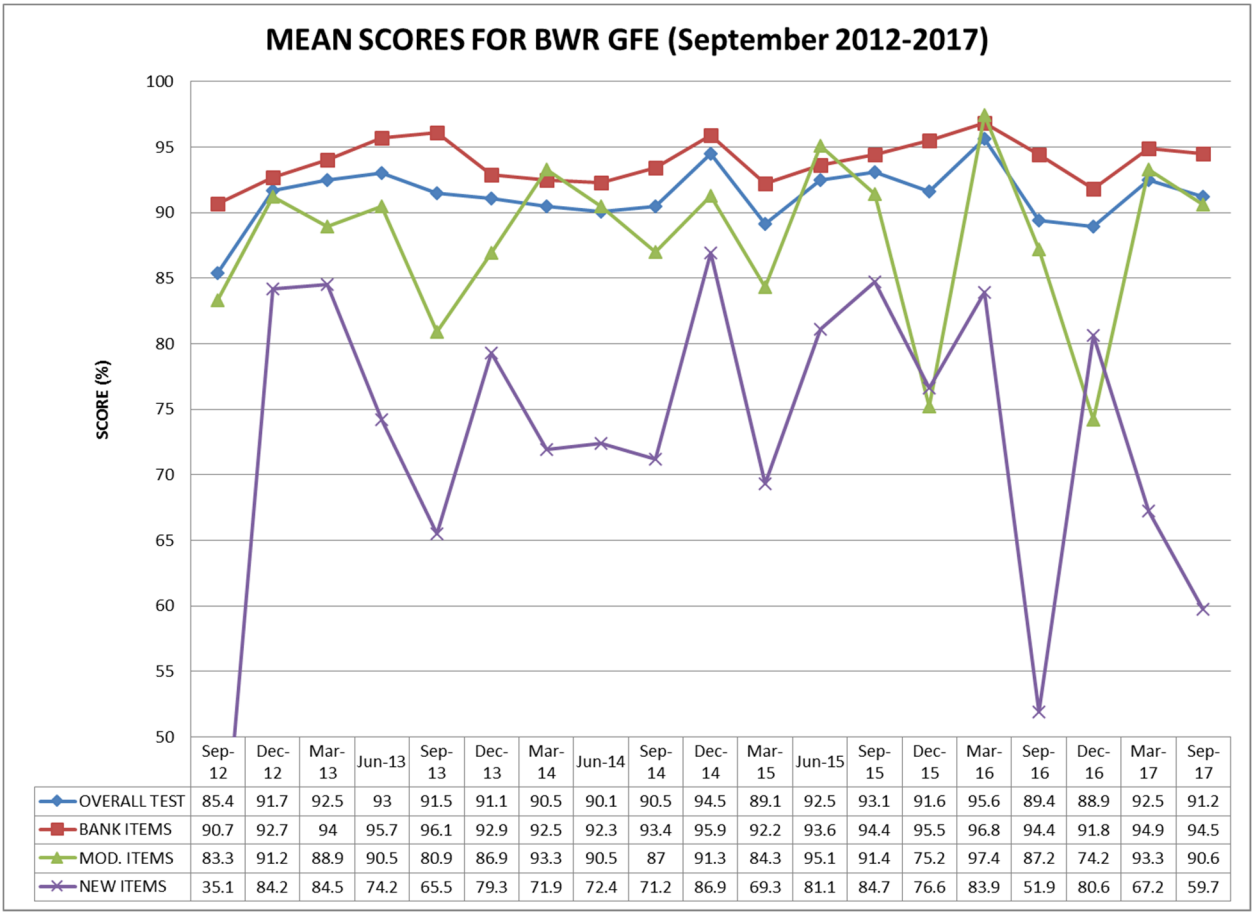
These results suggest that the reused items, previously disclosed and available for test taker rehearsal and review, were easier and less discriminating that the modified or new test items. Dr. Usova concludes that, “the balanced use of test item banks with modified and new items combine to produce a discriminating exam.” Dr. Usova further states that, “test item banks serve as a valuable resource for learning and represent a resource for training and test development; however, when all or too high a portion of items for an examination are drawn from the validated [test] bank and are identical to those items that have been previously used for testing, the banks are inappropriately used.”

RES staff replicated Dr. Usova’s study by examining scores on the GFE within the most recent 5 year period. Figure 5 shows the mean scores on the PWR GFE from September 2012 through September 2017. Figure 6 shows the means scores during the same time period for the BWR GFE. The same pattern observed by Dr. Usova is still present in the more recent performance data from the GFE. A lower percentage of examinees answer new test items correctly as compared to bank or modified test items.



	Sep-12	Dec-12	Mar-13	Jun-13	Sep-13	Dec-13	Mar-14	Jun-14	Sep-14	Dec-14	Mar-15	Jun-15	Sep-15	Dec-15	Mar-16	Jun-16	Sep-16	Dec-16	Mar-17	Sep-17
OVERALL TEST	89.6	94.6	94	93.4	93.2	91.9	91.4	90.8	91.5	93.3	93.4	93.5	93.9	93.1	94.1	89.9	92.6	92.4	92.1	90
BANK ITEMS	94.9	95.8	95.3	95	97	94.1	94.1	94	92.6	94	94.7	94.1	94.2	96.2	95.1	92.7	96.6	95	95.2	93.4
MOD. ITEMS	90.4	93.8	92	88.4	88.1	91.9	85.6	87	85.5	91.3	93.9	94.2	92.1	77.4	97.8	89.8	95.9	81	79.7	88.8
NEW ITEMS	35.6	85.9	85.8	82.9	67.7	74.8	75.9	69.5	88.7	90.4	82.2	88.4	93.4	83.7	82.7	67.4	58.3	82.7	77	57.5

Figure 5. Mean Scores on the PWR GFE (September 2012-2017)



	Sep-12	Dec-12	Mar-13	Jun-13	Sep-13	Dec-13	Mar-14	Jun-14	Sep-14	Dec-14	Mar-15	Jun-15	Sep-15	Dec-15	Mar-16	Sep-16	Dec-16	Mar-17	Sep-17
OVERALL TEST	85.4	91.7	92.5	93	91.5	91.1	90.5	90.1	90.5	94.5	89.1	92.5	93.1	91.6	95.6	89.4	88.9	92.5	91.2
BANK ITEMS	90.7	92.7	94	95.7	96.1	92.9	92.5	92.3	93.4	95.9	92.2	93.6	94.4	95.5	96.8	94.4	91.8	94.9	94.5
MOD. ITEMS	83.3	91.2	88.9	90.5	80.9	86.9	93.3	90.5	87	91.3	84.3	95.1	91.4	75.2	97.4	87.2	74.2	93.3	90.6
NEW ITEMS	35.1	84.2	84.5	74.2	65.5	79.3	71.9	72.4	71.2	86.9	69.3	81.1	84.7	76.6	83.9	51.9	80.6	67.2	59.7

Figure 6. Mean Scores on the BWR GFE (September 2012-2017)

RES staff also extended the study by statistically testing these observed differences. The purpose of the statistical testing was to determine whether differences between mean scores in the sample were due to chance. The overall mean scores for new, modified, and bank items are presented in Table 2.

Table 2. Mean Scores by Item Type on the PWR and BWR GFEs from 2012-2017

Exam Type	Item Type	Item Statistics			
		Mean	N	Std. Deviation	Std. Error Mean
PWR GFE	New Item	76.53	20	14.00	3.13
	Modified Item	89.23	20	5.37	1.20
	Bank Item	94.70	20	1.16	0.26
BWR GFE	New Item	72.64	19	13.01	2.99
	Modified Item	88.03	19	6.20	1.42
	Bank Item	93.91	19	1.66	0.38

Note. The BWR GFE was not administered in June 2016, thus N=19 for the BWR GFE and N=20 for the PWR GFE.

A paired samples t-test was used to test the significance of the differences between the mean values for new, modified, and bank test items. As can be seen in Table 3, the mean differences were statistically significant, meaning that the differences between new, modified, and bank

items on the GFE are not likely due to chance. In practical terms, the statistical significance indicates that that observed differences are reliable and represent a true variation in performance based on item type.

Table 3. Test of Differences between Mean Scores for New, Modified, and Bank Items

Paired Samples T-Test										
			Paired Differences					t	df	Sig. (2-tailed)
			Mean Diff.	Std. Dev.	Std. Error Mean	95% Confidence Interval of the Difference				
						Lower	Upper			
PWR GFE	Pair 1	New - Bank	-18.18	14.17	3.168	-24.80	-11.55	-5.74	19	.000
	Pair 2	Modified - Bank	-5.47	5.50	1.23	-8.05	-2.89	-4.45	19	.000
	Pair 3	New - Modified	-12.71	15.17	3.39	-19.80	-5.61	-3.75	19	.001
BWR GFE	Pair 1	New - Bank	-21.27	12.57	2.88	-27.33	-15.21	-7.37	18	.000
	Pair 2	Modified - Bank	-5.88	6.01	1.38	-8.78	-2.99	-4.27	18	.000
	Pair 3	New - Modified	-15.38	13.09	3.00	-21.69	-9.08	-5.12	18	.000

This research on scores by item type suggests that GFE examinees are significantly more likely to answer question bank questions correctly as compared to new or modified questions. Thus, there is evidence that recall of question bank items has an effect on GFE performance. In addition, the observed effect continues to be present in more recent data when the question bank has substantially increased in size.

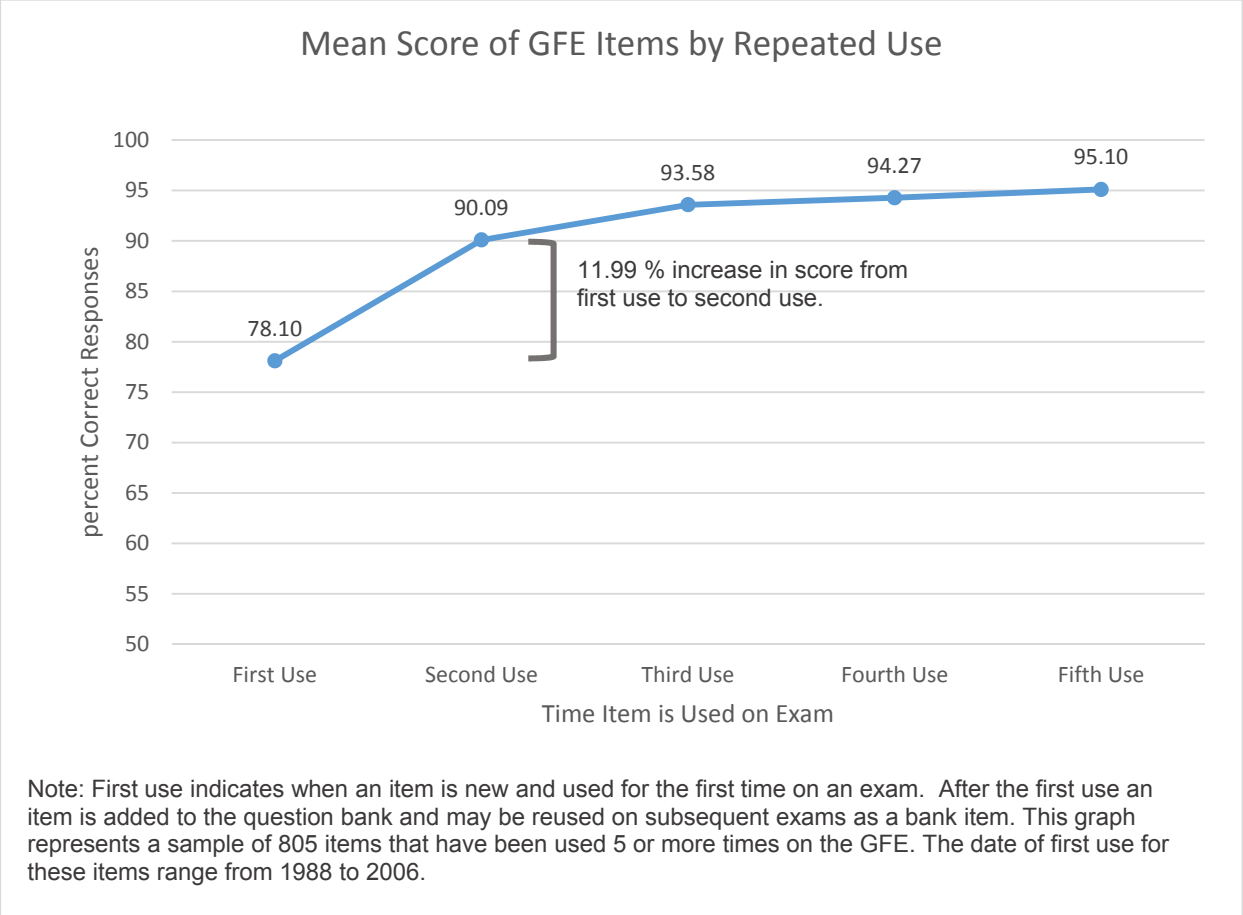
4.7. Research Comparing GFE Item Difficulty Level with Repeated Use

One of the potential limitations of the study conducted by Dr. Usova is that differences between the mean scores for new and bank items may be due, in part, to systematic differences in the characteristics of the items themselves. It is possible that examinees score lower on new items because, for example, the content of the new item is more difficult, the structure or wording of the item is confusing, or the item is not as familiar to examinees because it is dissimilar from items in the question bank. Dr. Usova speculates that, “the reason for the lower performance of new items is likely attributable to their newness alone since there is no intended basis, during their development, to make them inherently any more difficult than the remaining items” (p. 107). In addition, the guidance for written examinations in NUREG 1021, Appendix B recommends avoiding use of questions that are unnecessarily difficult (p. B-5). As a result, new questions that fall well above or below the recommended difficultly level of 70 to 90 percent may not be used on future examinations because they do not appropriately discriminate among examinees.

One way to control for potential systematic differences between new and bank items is to examine the difficulty level of items that are used on multiple examinations. The first time an item is used on a GFE it is considered a “new item.” If the new item meets specified psychometric criteria (e.g., appropriate difficulty level, discrimination ability, etc.), then it is added to the GFE question bank and becomes available for use on future GFEs as a “bank item.” It is expected that the psychometric properties of an item should not appreciably change each time it is used on an exam. The primary difference between an item the first time it is used on a GFE as a “new item” and subsequent times it is reused on a GFE as a “bank item” is that the item has been made available to examinees for review and study in the GFE question bank. Therefore, any observed differences in performance of an item the first time it is used as

compared to subsequent times it is reused are much more likely to be attributable to the effects of disclosing that item to examinees in the GFE question bank.

RES staff examined the historical data for items that have been used on multiple GFEs to determine if there was a difference in mean scores (i.e., percent of examinees who answered the item correctly) the first time an item is used as compared to subsequent uses after the item has been added to the question bank. Items currently listed in the PWR and BWR question banks that had been used without modification on five or more GFEs were selected for inclusion in the sample. Overall, the sample consisted of 805 items. RES staff then calculated the mean score for each item the first time it was used on a GFE (i.e., as a new item) and subsequent times when it was used on a GFE (i.e., as a bank item).



Note: First use indicates when an item is new and used for the first time on an exam. After the first use an item is added to the question bank and may be reused on subsequent exams as a bank item. This graph represents a sample of 805 items that have been used 5 or more times on the GFE. The date of first use for these items range from 1988 to 2006.

Figure 7. Mean Scores of GFE Items by Repeated Use

As can be seen in Figure 7, the mean percent of correct responses increased with each use of the item. The mean score when an item was first used was 78.10 percent. The mean score then increased to 90.09 percent the second time an item was used on the GFE. This indicates that the overall difficulty level of the item decreases as the item is used repeatedly on an exam.

RES staff tested whether the differences between the means were statistically significant using paired-sample t-tests (Table 4). In each case, the differences between the paired means were statistically significant. For instance, the mean score for items the first time they were used on the GFE was significantly different from the mean score the second time an item was used.

Further, the mean score the second time an item was used was significantly lower than the mean score the third time the item was used. The statistical significance of this effect indicates that the observed decrease in difficulty level is reliable and is not likely to be due to chance.

Table 4. Test of Differences between Mean Scores when an Item is New (First Use) and Subsequently Reused

Paired Samples T-Test									
		Paired Differences					t	df	Sig. (2-tailed)
		Mean Diff.	Std. Dev.	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	First Use – Second Use	-11.99	16.93	.60	-13.16	-10.82	-20.09	804	.000
Pair 2	Second Use – Third Use	-3.49	10.84	.38	-4.24	-2.74	-9.15	804	.000
Pair 3	Third Use – Fourth Use	-.69	8.07	.28	-1.25	-.13	-2.43	804	.015
Pair 4	Fourth Use – Fifth Use	-.82	7.85	.28	-1.37	-.28	-2.98	804	.003

Because each item is identical over repeated use, it is highly likely that the nearly 12 percent increase in score (decrease in difficulty) from first use to second use is due to the item's disclosure within the question bank. The mean score also continues to increase with repeated use on an exam. When an item is used for the fourth or fifth time, it has been available for study within the GFE question bank for a significant period of time. The likelihood that an examinee will answer a question correctly substantially increases when an item is disclosed in the question bank, and continues to increase the longer the item is exposed and available for study.

4.8. Summary on Question Bank Size and Bank-Only Exams

Although generally, as a question bank increases in size, the likelihood that examinees would be able to memorize and recall the entire bank decreases, there is no research basis to support drawing a conclusion about how large the question bank would need to be to support a question bank-only GFE. The Hale study (1980) found a smaller effect on performance when a question bank was larger in size (1800 items as compared to 900 items). However, in the case of the GFE, there continues to be a statistically significant difference in performance between new items and bank items even though the question bank has nearly doubled in size since 1996. Further, overlap among questions in the question bank can make it easier to memorize and recall a large number of questions, and the public availability of the question bank provides examinees with ample time to study.

The current research also demonstrates that differences in performance between new and bank items in the GFE are not likely to be because new questions are inherently more difficult—the same questions, when added to the question bank and reused on subsequent exams had an approximately 12% decrease in difficulty. Note that with the current use of bank, modified, and new items on the GFE, the exam developers are able to implicitly correct for a bank item's decrease in difficulty after first use by 1) introducing new and modified items to the exam, and 2) using a balanced mix of more and less difficult items, to ensure that the overall difficulty level of the exam remains relatively stable. Based on this research, removing new and modified items from the exam would likely decrease the overall difficulty level of each exam. As a result, new examinees would face a significantly less difficult exam than previous examinees, and the reliability and validity of the exam would be compromised.

5. Other Options for Test Item Development

Rather than relying solely on the question bank for GFE development, there are other options that may be explored to increase the flexibility of the GFE while controlling administration costs. The following section provides some ideas for consideration.

5.1. Limit Item Disclosure

One approach that could be considered is to begin populating a separate bank of new and modified items that are not disclosed to the public. Items could be retained in the undisclosed item bank for a set period of time, or until they are no longer used on the GFE. Future exams could then be developed by taking 40 questions from the disclosed item bank, 5 questions from the undisclosed modified item bank, and 5 questions from the undisclosed new item bank. If adequate numbers of questions are included in the undisclosed item banks, then those questions could be sampled from repeatedly prior to being disclosed to the public. However, controls would need to be established regarding repeat test-takers, to ensure that a test-taker was not exposed to the same new or modified test questions due to repeated attempts at taking the exam. In addition, the undisclosed items should be carefully secured and monitored while in use to ensure they are not disclosed to examinees, whether intentionally or unintentionally. Consistent with guidance for linear-on-the-fly testing, the undisclosed item bank should contain at least 10 times the number of items needed for any one exam (Prometric, 2017).

5.2. Suggestion Box for New Items

Another option for increasing the number of new items available for use on exams would be to solicit suggestions for questions from utilities and the public. For example, the FAA website includes a link where any member of the public can submit new questions for airman knowledge examinations. The link includes guidance for submitting questions to ensure that the submission indicates the content domain from the airman certification standards and that the submission is formatted in a manner consistent with other questions on the examination (e.g., four option multiple-choice format). Accepting suggestions for questions could be a low-cost option for generating new content. The GFE developers could review the suggestions for suitability and modify questions as needed to fit the accepted format as outlined in NUREG-1021.

5.3. Item Cloning and Item Modeling

An option for increasing the generation of modified items would be to explore use of item cloning or item modeling. Item cloning involves making minor changes to an item such that the content domain represented by the item is intact, but requires a different correct response (Fitzgerald & Mulkey, 2013). Multiple items can be created with slight variations. Excluding the cloned items from the publicly disclosed question bank may increase the security of the test while reducing long-term item development costs. Item modeling involves assembling new items based on a standard template, which includes the item stem, options, and other auxiliary information (Gierl, Zhau, Alves, 2008). The stem forms the content and context for the question the examinee is required to answer, the options contain the possible answers, and the auxiliary information includes any additional text, figures, or other graphics required to answer the question. By systematically manipulating the stem, options, and auxiliary information, large numbers of different items can be created from each item model. However, pilot testing would be necessary to determine whether items developed through cloning or modeling achieve an acceptable level of item difficulty and do not substantially increase the predictability of the exam.

6. Computer-Based Testing

The GFE has been administered as a standard paper-and-pencil test since it began in 1988. As computer technology has developed, inquires have been made regarding the feasibility of transitioning the GFE to a computer-based format. This section presents various types of computer-based testing, research comparing computer-based testing to standard paper-and-pencil testing, and additional considerations when developing a computer-based administration strategy.

6.1. Types of Computer-Based Tests

There are many different types of computer-based tests. Computer-based tests can vary in terms of their underlying theory, test delivery model, and level of complexity. One means of distinguishing between different types of computer-based tests is based on the extent to which the test is adaptive to an examinee's performance during the exam (College Board, 2000).

At one end of the continuum are linear tests, which are fixed and do not change as a result of an examinee's performance. Linear tests have the same properties as paper-and-pencil tests, but are administered on the computer. Linear tests are based on classical test theory and are the most common type of standardized test. In linear testing, identical forms are given to examinees, and examinees are allowed to review, revise, and skip items as they proceed through the test. These types of tests are developed in the same manner as paper-and-pencil tests and are, therefore, easier to implement and more familiar to examinees.

At the other end of the continuum are fully adaptive tests, where the presentation of each item in the test is based on the examinee's performance on previous items. Adaptive tests use item response theory to estimate an examinee's proficiency based on his or her performance on a set of items. Adaptive tests are often used to classify an examinee's performance relative to other examinees. For example, the Graduate Records Examination (GRE) is an adaptive test used for admission to graduate and business schools in the United States. The GRE provides a standardized measure for comparing applicants' qualifications based on verbal reasoning, quantitative reasoning, and analytical writing. One strength of an adaptive test is that it can differentiate between levels of mastery of a subject area. Classifying levels of mastery may not be necessary or relevant for screening exams like the GFE where pass/fail decisions are made based on a pre-determined cut score. Adaptive tests can also be relatively complex to implement because they require significant resources to calculate measurement parameters for each item, specify item selection procedures, and develop large item banks. The complexity of adaptive tests makes them less cost effective for low-volume exams like the GFE, which is typically administered to less than 500 individuals each year.

Other types of test delivery models include linear-on-the-fly testing and use of testlets (College Board, 2000). Linear-on-the-fly testing involves developing unique, fixed-length tests for each examinee at the beginning of a test session. Each unique form is assembled based on pre-defined content and psychometric specifications. Linear-on-the-fly testing is similar to linear testing, but it allows for randomized tests to be developed on-demand. Another variation of computer-based testing is the use of testlets. A testlet is a group of items that are considered a unit and administered together. Each testlet is developed in advance and represents a unique group of items. Testlets may be constructed based on content specifications, difficulty level, or other psychometric characteristics of the items. Multiple testlets can then be randomly selected at the beginning of a testing session for on-demand testing. Testlets can also be used in

adaptive testing, with more or less difficult testlets selected at different stages of the test based on how an examinee performed on previous testlets.

6.2. Computer-Based vs. Paper-and-Pencil Testing

Research examining differences between computer-based and paper-and-pencil testing has produced mixed results (Mead & Drasgow, 1993). When differences are found between computer-based and paper-and-pencil testing, it is most often attributed to differences in presentation of the test (e.g., need to scroll or move between pages on a computer screen) and comfort level with the process of taking a test on a computerized interface (Poggio et al., 2005). The more similar the presentation of the computer-based test is to the original paper-and-pencil test, the less likely that the validity and reliability of the test will be affected by the different mode of administration. For instance, presenting each test question in its entirety on a single screen, allowing test-takers to skip questions and return to questions at a later time, and ensuring the process of selecting responses (e.g., with mouse clicks) is easy to understand and use. Test-takers may also be allowed to use scratch paper to make notes when answering test questions, so long as the scratch paper is not removed from the test center.

Use of computers has progressed rapidly over the past three decades, such that computer use is now very common in most work environments. Most adult workers in the nuclear industry can be expected to be familiar with using standard computer applications and executing basic commands (e.g., use of a keyboard and computer mouse). Thus, it is unlikely that lack of familiarity with computers will be a construct-contaminating factor in the exam (see discussion in Section 2.1). In addition, computer technology has advanced to the point that concerns about presentation of test questions on a computer screen can be mitigated. In some cases, the computer screen may offer usability advantages over traditional paper-based tests because computer monitors are commonly larger than a traditionally-sized piece of paper, and also test-takers may be able to zoom in and out when viewing complex diagrams.

The primary advantage of using computer-based testing over paper-and-pencil testing is the ability to leverage the electronic format of the test. Administration costs may be reduced by reducing or eliminating the need to print, mail, and return paper tests. Scoring of the test may be performed automatically in real time when the test is completed. Data from the test may be imported into database management systems to track and maintain statistics on the performance of examinees and the psychometrics of individual test items. Test security may also be strengthened by reducing the availability of copies of the test prior to its administration.

There are also a number of new considerations when moving from paper-and-pencil testing to computer-based testing. For instance, steps must be taken to ensure the security of the computers used for testing and the databases where testing information is stored. The presentation of test questions on the computer screen should be pilot tested to ensure readability and usability of the computer interface. Logistics such as securing test facilities with computers and determining how the exam will be uploaded to the computers will also needed to be considered.

6.3. Considerations for Computer-Based Test Administration

A number of different factors should be considered when choosing an approach to administering a computer-based test. Maintaining a database of psychometrics on item difficulty, content, and exposure is important to ensure each exam is developed in a manner that is consistent with the

construct(s) that the exam is intended to measure. At a minimum, the item database should include the following information about item characteristics:

- Item Number – unique label for tracking purposes.
- Item Difficulty – proportion of examinees who answered the question correctly.
- Item Discrimination –correlation between examinee responses to a particular item and responses on all other items on the test. A higher correlation indicates that examinees who answered the question correctly tended to do better on the exam and examinees who answered the question incorrectly tended to do worse on the exam. Thus, items with high correlations are better at discriminating between good and poor performers on the exam.
- Item Exposure – number of times the item has been used on past exams. Additional information on exposure may include specific administration dates when an item was used or whether an item appears in the disclosed question bank.
- Knowledge and Ability Catalog Number – indicates how the item is related to the content domains that the test is intended to sample.
- Relationships with Other Items (e.g., modified from question X) – indicates whether an item was developed as a modified version of another item or whether an item has similar content to other items in the question bank. This information can be used to ensure that items selected for the same exam are not interdependent and do not provide cues that can be used to answer other items.

With computer-based testing, algorithms can be created to automate the selection of items for inclusion on an exam based on their psychometric characteristics. If linear computer-based testing is used, subject-matter experts can review the exam prior to its administration to ensure that the automated process of selecting items has created an appropriately discriminating exam. Over time, refinement of the test items and algorithms used to select items may facilitate a transition to linear-on-the-fly testing, where unique exams can be created on-demand. Subject-matter experts may then transition from reviewing each exam to reviewing samples of exams based on the item selection algorithms. However, the importance of item security increases as more complex computer-based testing modes are used and there are fewer opportunities for subject-matter expert review of individual exams.

7. Conclusions

The RES assessment concludes that GFEs derived solely from items in the GFE question bank would decrease the validity and reliability of the exam and cannot assure examination integrity. This conclusion is consistent with the technical basis and recommendations on test development from NUREG-1021. Given that the GFE question banks are and have been disclosed to the public, the size of the bank should not be used as a justification for deriving all examination items from the question bank.

The impacts on the validity and reliability of the exam include:

- **Decrease in cognitive level of exam questions:** When an exam is derived from items in the question bank, the correct answers and distractors do not change, which increases the probability that examinees may select the correct answer because they recognize the question from their studies, rather than because they comprehend the material. A higher proportion of test items could be answered correctly based on recall, rather than through problem-solving, comprehension and analysis. A bank-only examination would no longer involve testing higher cognitive response processes.
- **Increase in predictability of exam:** The GFE is a high-stakes test where examinees are highly motivated to pass. Although it is unlikely that most examinees would be able to memorize and accurately recall all of the items in the question bank, it is reasonable to expect that deriving all questions from the question bank would encourage examinees to study previous tests at the expense of studying the entire content domain from which the test is intended to sample. A bank-only examination would, therefore, violate the validity inference that performance on the GFE is an accurate assessment of the fundamental knowledge required to safely and competently operate a nuclear power plant.
- **Decrease in difficulty:** The RES investigation of GFE data revealed that items from the question bank are significantly less difficult than new and modified items. Further, new items showed a decrease in difficulty of nearly 12 percent when they were reused on subsequent exams. Statistical analysis shows this decrease is significant and not due to chance. These results suggest that the likelihood that an examinee will answer a question correctly substantially increases when an item is disclosed in the question bank, and continues to increase the longer the item is in the question bank and available for study. A bank-only examination would likely be less difficult than exams with new and modified questions.
- **Decrease in discrimination ability:** Each GFE is implicitly developed by subject-matter experts to ensure that the exam questions are a representative sample of the content domains, engage higher cognitive response processes, and are at an appropriate level of difficulty to discriminate among examinees who have and have not mastered the underlying fundamentals of nuclear power reactor operations. The 80 percent cut-off score may no longer be a valid criterion for discriminating among examinees who take a bank-only exam because the test questions are more predictable, less difficult, and no longer require higher cognitive response processes. As a result, a bank-only exam may disproportionately benefit examinees who have not mastered the material.

However, reliance on the question bank is not the only option for increasing examination flexibility and lowering administration costs. RES staff concludes that linear computer-based

testing is a viable alternative to the paper-based GFE. Linear computer-based testing can be implemented in a way that is very similar to a standard paper-based test, and thereby minimize the likelihood that changing the mode of administration will affect the validity and reliability of the exam. Although there may be up-front costs associated with transitioning to computer-based testing, cost-savings may be realized over time by making the process of developing and administering the exam more efficient. Further, exclusive use of the disclosed question bank is not a necessary prerequisite for transitioning to computer-based testing. Other options also exist for developing new items and ensuring item security to protect the integrity of the exam. Transitioning from paper-based to linear computer-based testing may not require significant changes to the process currently employed to develop the GFE. Instead, the primary change would be limited to the mode of administration (i.e., computer vs. paper). Once a computerized process for administering the exam is developed, it may be more feasible to explore more significant process changes in the future, such as linear-on-the-fly testing.

The continued inclusion of new and modified items remains critical to the ongoing validity of the exam. New and modified items ensure that the exam is not entirely predictable. Even if some test items seem familiar to examinees they must still engage higher cognitive levels (i.e., comprehension and analysis) to confirm the correct answer. Examinees are also forced to study content material beyond what is included in the question bank, thereby supporting the validity inference that passing the exam indicates sufficient mastery of the underlying knowledge fundamentals. This balanced approach leverages the benefits of using previously developed items from the question bank (i.e., saves costs associated with time and effort to develop new items and increases transparency of the examination) while still ensuring the exam is not overly predictable and remains capable of discriminating between examinees who have and have not mastered the fundamentals tested on the exam.

8. References

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

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., Wittrock, M.C. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives*. New York: Pearson, Allyn & Bacon.

Bloom, B.S. (Ed.). Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain.* New York: David McKay Co Inc

Burghof, K. L. (2001). Assembling an item-bank for computerised linear and adaptive testing in Geography. *Proceedings of the Educational Research Conference Special Issue: Measuring Cognitive and Non-Cognitive Systems of Reasoning: Some Preliminary Findings,* 2(4), 74.

Casner, S. M., Jones, K. M., Puentes, A., & Irani, H. (2004). *FAA Pilot Knowledge Tests: Learning or Rote Memorization? (NASA/TM—2004–212814).* National Aeronautics and Space Administration, Ames Research Center, Moffett Field, California.

College Board. (2000). *An Overview of Computer-Based Testing.* Research Note 09. Office of Research and Development, New York, NY. Retrieved November 9, 2017 from https://research.collegeboard.org/publications/content/2012/05/overview-computer-based-testing.

Federal Aviation Administration. (2017). *Airman Testing Standards Branch Questions and Answers.* Retrieved November 7, 2017 from https://www.faa.gov/training_testing/testing/media/questions_answers.pdf

Fitzgerald, C. T., & Mulkey, J. R. (2013). Security planning, training, and monitoring. *Handbook of Test Security,* 127-146.

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning and Assessment*, 7(2).

Hale, G. A., Angelis, P. J., & Thibodeau, L. A. (1980). *Effects of Item Disclosure on TOEFL Performance.* ETS Research Report Series No. 8. Educational Testing Service, Princeton, NJ.

National Research Council. (1991). *Performance Assessment for the Workplace, Volume 1.* The National Academies Press, Washington, DC. https://doi.org/10.17226/1862.

Nuclear Regulatory Commission. (2014). *Operator Licensing Examination Standards for Power Reactors (NUREG-1021, Revision 10).* Nuclear Regulatory Commission, Washington, DC.

Nuclear Regulatory Commission. (2017). *Knowledge and Abilities Catalog for Nuclear Power Plant Operators: Pressurized Water Reactors (NUREG-1122, Revision 3 Draft for Comment).* Nuclear Regulatory Commission, Washington, DC.

Nuclear Regulatory Commission. (2017). *Knowledge and Abilities Catalog for Nuclear Power Plant Operators: Boiling Water Reactors (NUREG-1123, Revision 3 Draft for Comment).* Nuclear Regulatory Commission, Washington, DC.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment*, 3(6). Retrieved November 9, 2017 from http://www.jtla.org

Prometric. (2017). *Exam Bank Development: How to Build in Security and Flexibility.* Retrieved November 17, 2017 from https://www.prometric.com/en-us/news-and-resources/reference-materials/pages/Exam-Bank-Development.aspx.

Roth, E. M., Mumaw, R. J., & Lewis, P. M. (1994). *An Empirical Investigation of Operator Performance in Cognitively Demanding Simulated Emergencies (NUREG/CR--6208).* Nuclear Regulatory Commission, Washington, DC. (United States).

Usova, G. M. (1997). Effective test item discrimination using Bloom's taxonomy. *Education, 118(1)*, 100-110.