Final report of the conduct of a feasibility study for the computerization and implementation of a uniform CPA examination in fifty-four jurisdictions

American Institute of Certified Public Accountants. Joint AICPA/NASBA Computerization Implementation Committee

Follow this and additional works at: https://egrove.olemiss.edu/aicpa_exam
Part of the Accounting Commons, and the Taxation Commons

Recommended Citation
https://egrove.olemiss.edu/aicpa_exam/169

This Book is brought to you for free and open access by the American Institute of Certified Public Accountants (AICPA) Historical Collection at eGrove. It has been accepted for inclusion in Examinations and Study by an authorized administrator of eGrove. For more information, please contact egrove@olemiss.edu.
FINAL REPORT OF
THE CONDUCT OF A FEASIBILITY STUDY
FOR THE COMPUTERIZATION AND IMPLEMENTATION
OF A UNIFORM CPA EXAMINATION IN FIFTY-FOUR JURISDICTIONS

PREPARED FOR THE
JOINT AICPA/NASBA COMPUTERIZATION IMPLEMENTATION COMMITTEE

AMERICAN INSTITUTE OF CERTIFIED PUBLIC ACCOUNTANTS
HARBORSIDE FINANCIAL CENTER
201 PLAZA III
JERSEY CITY, NJ 07311-3881

PREPARED BY
PROFESSIONAL EXAMINATION SERVICE
475 RIVERSIDE DRIVE
NEW YORK, NY 10115-0089

SEPTEMBER 1999
## JOINT AICPA/NASBA COMPUTER IMPLEMENTATION COMMITTEE

<table>
<thead>
<tr>
<th>Name</th>
<th>Company/Position</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>William W. Holder, CPA, DBA, Chair</td>
<td>University of Southern California</td>
<td>Los Angeles, CA</td>
</tr>
<tr>
<td>David L. Landsittel, CPA</td>
<td>Arthur Andersen LLP (Retired)</td>
<td>Winnetka, IL</td>
</tr>
<tr>
<td>Eric L. Schindler, CPA</td>
<td>American Chemet Corporation</td>
<td>East Helena, MT</td>
</tr>
<tr>
<td>Asa L. Hord, CPA</td>
<td>Deloitte &amp; Touch LLP (Retired)</td>
<td>Louisville, KY</td>
</tr>
<tr>
<td>Carol Sigmann</td>
<td>California Board of Accountancy</td>
<td>Sacramento, CA</td>
</tr>
<tr>
<td>Dennis P. Spackman, CPA</td>
<td>Church of Jesus Christ of Latter-day Saints</td>
<td>Salt Lake City, UT</td>
</tr>
</tbody>
</table>

## AICPA/NASBA STAFF

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anat Kendal, CPA</td>
<td>AICPA Director of Examination Reformation and Computerization</td>
</tr>
<tr>
<td>Arleen R. Thomas, CPA</td>
<td>AICPA Vice President Professional Standards and Services</td>
</tr>
<tr>
<td>James D. Blum, CPA, PhD</td>
<td>AICPA Director of Examinations</td>
</tr>
<tr>
<td>Craig Mills, EdD</td>
<td>AICPA Executive Director of Examinations</td>
</tr>
<tr>
<td>Lorraine P. Sachs, CAE</td>
<td>NASBA Executive Vice President</td>
</tr>
</tbody>
</table>
PREFACE

On behalf of Professional Examination Service (PES), we are pleased to have conducted this very important feasibility study for the American Institute of Certified Public Accountants (AICPA) regarding the computerization and implementation of the Uniform CPA Examination. This report addressed four sets of issues with regard to the proposed ideal examination:

- Time frames for computerizing the ideal examination, including the feasibility of computerizing and administering the Day 1 examination, using a computer mastery testing (CMT) model, and the Day 2 examination, comprising four simulations including free-response questions, by Spring 2003
- Financial viability of computerizing the ideal examination, including costs to develop, computerize, and administer the examination
- Assessment of the technical viability of the computerized ideal examination, including psychometric evaluation of the CMT model and the simulation component of the ideal examination
- Assessment of stakeholders’ reactions to the computerized ideal examination, including candidates’, boards’, and practitioners’ acceptance

This report presented conclusions in regard to the feasibility of implementing the ideal examination, recommendations regarding required activities for implementation, and a proposed method for making a final decision regarding the testing model(s) to be implemented.

A project of this magnitude depends on the hard work and commitment of many individuals, and we are pleased to acknowledge their contributions to the final product. This report represents a substantial investment of the AICPA’s resources. We appreciate the AICPA’s support and endorse the efforts of the AICPA Board of Examiners, Chaired by David B. Pearson, to develop an exemplary Uniform CPA Examination administered on the computer.

We are grateful to the Joint AICPA/NASBA Computerization Implementation Committee (CIC) for the advice and direction it provided. Its six members, William W. Holder (Chair), Asa L. Hord, David L. Landsittel, Carol Sigmann, Eric L. Schindler, and Dennis P. Spackman, participated in interviews and meetings, and regularly provided thoughtful feedback regarding reports and presentations.

We recognize the substantial contributions of the many experts who gave generously of their time and resources. To conduct the study, data were gathered from experts representing the AICPA, the National Association of State Boards of Accountancy, and individual state boards; and from external sources including psychometricians, computer-based testing vendors, test developers, test administrators, representatives of various credentialing agencies currently developing or implementing CBT programs, and other professionals familiar with issues relevant to the development and implementation of the AICPA’s ideal examination. Quite simply, this report would not have been possible without their input and advice.
In addition, we wish to acknowledge the contributions of Ronald K. Hambleton, the AICPA’s consulting psychometrician, and Michael T. Kane, NASBA’s external consultant, who reviewed and provided feedback regarding an alternate method for making a final decision about the testing model(s) to be implemented. We wish to thank the members of the AICPA and NASBA staff associated with the project, including: Arleen R. Thomas, AICPA Vice President Professional Standards and Services, and James D. Blum, AICPA Director of Examinations, for input regarding budgetary issues; Lorraine P. Sachs, NASBA Executive Vice President for insight into administration-related issues; and Craig Mills, AICPA Executive Director of Examinations, for conceptual guidance and leadership throughout the course of the study. Our primary contact for the study, Anat Kendal, AICPA Director of Examination Reformation and Computerization, made our work much easier than it might have been by providing key direction throughout the course of the study.

Complex studies such as this require solid internal and external logistical support. Janice D. Scheuneman, PES’s external consultant, was responsible for the data gathering and synthesizing initiatives, and Jeffery P. Mohn, PES Controller, prepared the cost estimates for the feasibility study.

We conclude by stating that the views expressed in this report are those of PES and do not necessarily reflect the views of the AICPA or NASBA, or of those experts who provided advice during the course of this feasibility study.

Sandra Greenberg, Vice President for Public Service Activities & Director of Research Programs
J. Patrick Jones, Executive Vice President
Ilsa Halpern, Director, Professional Development and Client Services

New York
September 1999
# Table of Contents

PREFACE .................................................................................................................. Page i
Table of Contents ....................................................................................................... Page iii

EXECUTIVE SUMMARY ................................................................................................ Page 1
Conclusions ................................................................................................................. Page 2
Feasibility in Regard to Time Frames, Cost ............................................................... Page 2
The time frames for computerizing the ideal examination ........................................ Page 2
The costs of developing and implementing the ideal examination ......................... Page 3
Required activities for implementation ..................................................................... Page 3
Technical and Stakeholder Concerns ....................................................................... Page 4
Day 1: Development and implementation of the CMT model ................................ Page 4
Day 2: Development and implementation of simulations ........................................ Page 4
Implementation of the comprehensive 2-day examination ...................................... Page 5
Recommendations .................................................................................................... Page 5
Establish financial parameters for initial examination development costs ............... Page 5
and ongoing operational costs ................................................................................. Page 5
Continue public outreach ......................................................................................... Page 5
Resolve potential problems ...................................................................................... Page 5
Make policy decisions regarding the information flow to candidates and other stakeholders ........................................................................................................ Page 5
Establish a method for making a decision regarding the testing model to be implemented ........................................................................................................ Page 5

OBJECTIVES OF THE FEASIBILITY STUDY ................................................................. Page 7

METHODOLOGY FOR THE CONDUCT OF THE FEASIBILITY STUDY ..................... Page 9
Data Collection Initiatives ........................................................................................ Page 9
Clarification of Associated Facts and Assumptions .................................................. Page 10
General assumptions regarding the administration of Day 1 and Day 2 of the examination ........................................................................................................ Page 10
General assumptions regarding the scoring of the Day 1 and Day 2 examinations ........................................................................................................ Page 11
Specific facts and assumptions regarding Day 1 of the ideal examination ............. Page 12
Specific assumptions regarding Day 2 of the ideal examination ............................ Page 13
Specific assumptions regarding the development and administration of a linear examination ........................................................................................................ Page 13
Assumptions regarding costing ............................................................................... Page 13
Overall assumptions concerning the roles and responsibilities of the boards of accountancy ........................................................................................................ Page 14
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVALUATION OF THE IDEAL EXAMINATION</td>
<td>15</td>
</tr>
<tr>
<td>Day 1: Computer Mastery Test</td>
<td>15</td>
</tr>
<tr>
<td>Technical Perspective on Day 1, CMT</td>
<td>16</td>
</tr>
<tr>
<td>Diagnostic subscores for “morning only” candidates</td>
<td>16</td>
</tr>
<tr>
<td>Content coverage for “morning only” candidates</td>
<td>16</td>
</tr>
<tr>
<td>Content coverage</td>
<td>17</td>
</tr>
<tr>
<td>Item exposure and test security</td>
<td>17</td>
</tr>
<tr>
<td>Testlet and item discrimination</td>
<td>17</td>
</tr>
<tr>
<td>Indeterminacy of passing status</td>
<td>17</td>
</tr>
<tr>
<td>Standard setting</td>
<td>18</td>
</tr>
<tr>
<td>Item calibration</td>
<td>19</td>
</tr>
<tr>
<td>Stakeholder Perspective on Day 1, CMT</td>
<td>19</td>
</tr>
<tr>
<td>Use of a CMT model and variable-length testing sessions</td>
<td>19</td>
</tr>
<tr>
<td>Content coverage</td>
<td>19</td>
</tr>
<tr>
<td>Timing: The Feasibility of Implementing Day 1 by 2003</td>
<td>20</td>
</tr>
<tr>
<td>Day 2: Simulation Test</td>
<td>20</td>
</tr>
<tr>
<td>Technical Perspective on Day 2, Simulations</td>
<td>21</td>
</tr>
<tr>
<td>Content coverage</td>
<td>21</td>
</tr>
<tr>
<td>Content validation studies</td>
<td>21</td>
</tr>
<tr>
<td>Scoring</td>
<td>22</td>
</tr>
<tr>
<td>Scoring of communications skills, including writing</td>
<td>22</td>
</tr>
<tr>
<td>Combining scores from different vendors</td>
<td>23</td>
</tr>
<tr>
<td>Reliability</td>
<td>23</td>
</tr>
<tr>
<td>Standard setting</td>
<td>23</td>
</tr>
<tr>
<td>Equating</td>
<td>24</td>
</tr>
<tr>
<td>Diagnostic feedback</td>
<td>24</td>
</tr>
<tr>
<td>Pretesting</td>
<td>25</td>
</tr>
<tr>
<td>Fidelity</td>
<td>25</td>
</tr>
<tr>
<td>General impact of using a limited number of simulations</td>
<td>25</td>
</tr>
<tr>
<td>Burden on the candidates</td>
<td>25</td>
</tr>
<tr>
<td>Security</td>
<td>25</td>
</tr>
<tr>
<td>Stakeholder Perspective on Day 2, Simulations</td>
<td>26</td>
</tr>
<tr>
<td>Response to Briefing Paper No. 1</td>
<td>26</td>
</tr>
<tr>
<td>The essay</td>
<td>26</td>
</tr>
<tr>
<td>Timing: The Feasibility of Implementing Day 2 by 2003</td>
<td>27</td>
</tr>
<tr>
<td>Implementation of a Comprehensive Two-Day Examination</td>
<td>27</td>
</tr>
<tr>
<td>The Conduct of a Practice Analysis</td>
<td>27</td>
</tr>
<tr>
<td>Evaluation of Content Coverage</td>
<td>27</td>
</tr>
<tr>
<td>Impact of Compensatory Scoring on an Examination Comprising Two Sections</td>
<td>28</td>
</tr>
<tr>
<td>Educational Materials Development</td>
<td>29</td>
</tr>
<tr>
<td>Implementation of a Centralized Examination Administration Call Center</td>
<td>29</td>
</tr>
<tr>
<td>Development of a Comprehensive Database and Software for Boards of Accountancy and Other Key Stakeholders</td>
<td>29</td>
</tr>
</tbody>
</table>
Costs ............................................................................................................................ Page 30
Development Costs ..................................................................................................... Page 30
  General development costs .................................................................................. Page 30
  Development of Day 1 examination ................................................................ Page 31
  Development of Day 2 examination ................................................................ Page 31
Implementation Costs ............................................................................................... Page 32
  Per candidate cost for Day 1 .............................................................................. Page 32
  Per candidate cost for Day 2 .............................................................................. Page 33
Benchmarking Costs for the Development of a Linear Examination ..................... Page 33
SUMMARY .................................................................................................................. Page 34
  Technical and Stakeholder Concerns ................................................................. Page 34
    Day 1: Development and implementation of the CMT model ....................... Page 34
    Day 2: Development and implementation of simulations ........................... Page 34
    Implementation of the comprehensive 2-day examination ......................... Page 34
  Feasibility in Regard to Time Frames, Costs ...................................................... Page 35
    The time frames for computerizing the ideal examination ......................... Page 35
    The costs of developing and implementing the ideal examination ............... Page 35
    Required activities for implementation ......................................................... Page 35
  Recommendations ............................................................................................... Page 36
    Establish financial parameters for initial examination development costs
      and ongoing operational costs ................................................................. Page 36
    Continue public outreach .............................................................................. Page 36
    Resolve potential problems ............................................................................ Page 36
    Make policy decisions regarding the information flow to candidates
      and other stakeholders .............................................................................. Page 36
    Establish a method for making a decision regarding the testing model
      to be implemented ................................................................................... Page 37
  Testing Models .................................................................................................... Page 37
  Testing Model Features for Comparison ............................................................ Page 38
Appendix 1: Psychometricians and Testing Construction Experts Interviewed
  and Discussion Topics .................................................................................... Page 45
Appendix 2: Exemplars of Related Research Literature ............................................ Page 48
Appendix 3: CBT Vendors .................................................................................... Page 56
Appendix 4: Credentialing Agencies and Information Technology Companies
  Implementing CBT Programs ........................................................................... Page 57
Appendix 5: Details Regarding the Facts and Assumptions .................................... Page 58
Appendix 6: Joint AICPA/NASBA CIC CBT Feasibility Study Cost Estimates ....... Page 66
EXECUTIVE SUMMARY

Professional Examination Service (PES) conducted a feasibility study on behalf of the American Institute of Certified Public Accountants (AICPA) regarding the computerization and administration of the Uniform CPA Examination. The form and content of the proposed computerized examination are based on independent research and input from various constituencies, and the results are currently identified as the "ideal exam." The proposed structure of the computer-based examination is a two-section examination. Each section would be taken and graded independently on a pass/fail basis. Testing for each section would be offered by appointment at regular intervals throughout the year in all jurisdictions. Candidates would be free to choose the order in which they take the sections. Diagnostic feedback would be provided to the candidates.

The main features of the ideal examination are:

- The first section would primarily test the basic knowledge needed by entry-level CPAs using multiple-choice questions. Day 1 is envisioned as a computer mastery test (CMT) wherein items are administered in groups known as testlets. Test length is variable, which means that some candidates are tested longer than others to determine if they pass or fail.

- The second section would emphasize assessing the skills needed by entry-level CPAs using performance assessment. Day 2 would consist of four simulations, about 1½ hours each in length, that would simulate real-world practice and assess higher-level, integrated knowledge and skills, including analytical, research, and communications skills.

The feasibility study addressed four sets of issues with regard to the ideal examination:

- Time frames for computerizing the ideal examination, including
  —the feasibility of computerizing and administering the Day 1 examination, using a computer mastery testing (CMT) model, by Spring 2003
  —the feasibility of computerizing and implementing the Day 2 examination, comprising four simulations including free-response questions, by Spring 2003

- Financial viability of computerizing the ideal examination, including
  —cost to develop an objectively scored item
  —cost to develop a testlet
  —cost to develop and computerize a simulation, including the cost of grading free-response questions
  —cost to computerize the ideal examination (including hardware/software)
  —cost to administer (deliver) Day 1 and Day 2 of the ideal examination

- Assessment of the technical viability of the computerized ideal examination, including
  —psychometric evaluation of the CMT model
  —psychometric evaluation of the simulation component of the ideal examination
• Assessment of stakeholders’ reactions to the computerized ideal examination, including candidates’, boards’, and practitioners’ acceptance

To conduct the feasibility study, data were gathered from subject-matter experts (SMEs) representing the AICPA, the National Association of State Boards of Accountancy (NASBA), and individual state boards; and from other sources including psychometricians, computer-based testing (CBT) vendors, test developers, test administrators, and other professionals familiar with issues relevant to the development and implementation of the AICPA’s ideal examination. PES reviewed research literature and AICPA, NASBA, and state board documents, and interviewed representatives of various credentialing agencies currently developing or implementing CBT programs.

Discussions with psychometricians and test developers focused on the following topics:

• Day 1 (CMT examination composed of objectively scored items): advantages and disadvantages of various testing models, e.g., CMT, linear, computer adaptive sequential testing (CAST), domain sampling, computer adaptive testing (CAT); length of testing session; procedures to develop and pilot-test items and assemble testlets and examinations; hardware and software requirements; potential for content coverage; reliability (and standard error of measurement); efficiency; size of item pool; generation of diagnostic reports; standard setting; candidates’ reactions and other stakeholders’ responses; security; cost.

• Day 2 (simulations): time needed to develop the program; procedures to develop the program; procedures to develop and pilot-test items and assemble simulations; general scoring issues (process versus content); potential for content coverage; relationship to objectively scored assessments; types of simulations (linear versus branched); level of interactivity; hardware and software requirements; standard setting; equating; generation of diagnostic reports; validity; candidates’ reactions and other stakeholders’ responses; security; cost.

Discussions with CBT vendors focused on administering the ideal examination to large numbers of candidates, the availability of secure networks, and the availability of hardware and software to support both Day 1 and Day 2 of the ideal examination. Discussions with representatives of credentialing agencies focused on candidates’ reactions and other stakeholders’ responses to the implementation of a CBT program, including the effects on credentialing agencies and licensing boards. Discussions with representatives of information technology companies provided insight into CBT in the business and industry sector.

Conclusions

Feasibility in Regard to Time Frames, Costs, and Implementation

The time frames for computerizing the ideal examination. Day 1 of the ideal examination can be implemented by 2003; however, the feasibility of implementing Day 2 by 2003 is somewhat
problematic. The outcomes of the AICPA's pilot project regarding the efficient development and scoring of valid and reliable simulations are required to properly evaluate this time frame.

The costs of developing and implementing the ideal examination. The ideal examination can be developed for $9.7 million, including the revision and implementation of the Content Specification Outlines (CSOs) subsequent to the conduct of the practice analysis; the development of educational materials and tutorials for candidates; the development of a single comprehensive database accessible to state boards and other key stakeholders; the development of items, forms, and simulations; computer programming for the development, administration, and scoring of the examination; the conduct of multiple cycles of reviewing and revising the examination; scoring; standard setting; equating; statistical and psychometric research and analysis; the selection of vendors; managing the program; and the implementation of quality control procedures throughout. The cost estimate for the ideal examination is based on the assumption that the results of AICPA's pilot project regarding the development and scoring of the simulations will provide support for the projected development and implementation schedule. Should additional cycles of examination development, review and revision be required, the cost will increase.

The ongoing per candidate cost for the ideal examination is estimated at $376, including $178 for Day 1 and $198 for Day 2. About two thirds of the total per candidate cost is attributable to costs for CBT vendor administration, e.g., “seat time” at a computer center. To estimate the CBT vendor administration cost, PES used the standards for secure delivery implemented by other high-stakes licensing examination programs administered on computer.

For purposes of comparing the cost of a CMT examination with the cost of a non-CMT examination, separate cost estimates were developed for a 1-day, multiple-choice, fixed-length linear examination, comparable in content coverage to Day 1 of the ideal examination. The per candidate cost for the linear examination is estimated at $179. Analysis and comparison of these virtually identical cost estimates for Day 1 of the ideal examination and for the linear examination indicate that the major cost driver for Day 1 is the number of hours of testing time rather than the underlying testing model.

Required activities for implementation. PES has identified several significant events that must occur in order to computerize and implement the CPA examination, regardless of the testing model(s) selected. First, the AICPA must contractually obligate one or more CBT vendors to provide a sufficient number of testing sessions for the candidates. Second, the AICPA must identify security issues associated with the delivery of the examination, and contractually obligate its CBT vendors to minimize the likelihood of security problems occurring. Third, the current update practice analysis related to General Business Knowledge and the full-scope practice analysis must be carefully monitored, and the results incorporated into revised CSOs, so that future objective-item and simulation development efforts reflect contemporary practice. Fourth, as the absolute number of objectively scored items and simulations required for CBT is greater than the number on hand, the AICPA must continue to “ramp up” its test development efforts to produce and pilot-test the quantities of items and simulations required for ongoing
implementation.

The AICPA must implement a research agenda to verify the facts and assumptions that PES made in regard to the ideal examination and/or any alternatives to that examination, because the accuracy of those facts and assumptions affected the conclusions reached regarding costs and timing. For example, the estimates PES made regarding the absolute number of candidates “in the pipeline” on an annual basis and the effects of using a compensatory scoring system for the Day 1 and Day 2 examination and/or a non-compensatory system for combining the Day 1 and Day 2 examination on the estimated pass/fail rates for either first-time or repeat candidates need to be verified through additional analyses. In addition, PES’s estimates of the frequency with which candidates might retake either day of the examination and the costs they would be willing to bear should be explored through further market research. Changes in the size of the candidate population will affect the numbers of items, testlets, forms, and simulations to be developed, pilot-tested, and published, in order to enhance examination security.

Technical and Stakeholder Concerns

**Day 1: Development and implementation of the CMT model.** Although a CMT testing model can be implemented, the use of a variable-length testing model has created problems for other high-stakes credentialing programs. Stakeholders, including both candidates and boards, may not understand CMT’s psychometric underpinnings, the fact that valid pass/fail decisions can be based on the administration of different numbers of items to different candidates, and may question the adequacy of the content coverage and the use of examination results for diagnostic feedback to these candidates. Because of the application of stopping rules, that is, rules that determine whether candidates have passed or failed the examination after the administration of only a base test and certain subsequent testlets, many candidates may fail a shortened test.

**Day 2: Development and implementation of simulations.** While simulations have excellent face validity because they closely mirror what practitioners do in their work, research is required to determine whether the scores they produce will be valid and reliable reflections of the knowledge and skills they are intended to measure.

PES identified several specific technical and psychometric concerns: content validity and the potential for inadequate coverage of the CSOs; unintentional effects that performance in one section may have on performance in another section of a simulation; the overall complexity of the simulations which require shifting among many computer screens and the use of on-line supplementary materials; the highly memorable content and associated security problems; the reliability of the simulations due to the limited number of scorable responses; feasibility of

---

1 In a compensatory model, pass/fail decisions are a function of total test scores which are additive across all sections, content areas, and/or subtests within the examination. Low scores on one or more elements in the examination are compensated for by high scores on other elements. In a non-compensatory model, pass/fail decisions are a function of test scores on each section, content area, and/or subtest. Low scores on one element are not offset by high scores on a different element.
setting the passing standard and equating; and difficulty in developing adequate diagnostic feedback to failing candidates.

Implementation of the comprehensive 2-day examination. While the Day 1 examination can produce valid and reliable scores that meet established psychometric criteria, the Day 2 examination, including only four 90-minute simulation problems, may not produce scores meeting defensible psychometric criteria in regard to reliability and validity. Furthermore, the content of the Day 1 and Day 2 sections may include either unintended overlap or omissions in content coverage.

The AICPA began its conduct of market research by issuing Briefing Paper No. 1—Conversion of the Uniform CPA Examination to a Computer-Based Examination, regarding the implementation of the ideal examination. Preliminary feedback from state board members indicates some initial misunderstanding regarding the description of the intended content and format of the ideal examination.

Recommendations

Establish financial parameters for initial examination development costs and ongoing operational costs. PES recommends that the Joint AICPA/NASBA Computerization Implementation Committee (CIC) investigate the amount of money that various stakeholders are willing to pay in order to finance the initial examination development costs and the ongoing operational costs of the examination.

Continue public outreach. PES recommends that communication strategies continue to be enhanced to inform key stakeholders of the progress toward computerization and implementation of the ideal examination. Communication initiatives need to be developed and implemented to obtain stakeholder views and information as well as stakeholder concurrence.

Resolve potential problems. PES notes the following potential problems that need to be investigated and/or resolved prior to implementation: jurisdictions’ and candidates’ reactions to the price of the examination; uniform conditioning requirements across the jurisdictions; uniform transition policies across the jurisdictions; validating the form and content of the assessment of communications skills; and validating the utility of Day 2 for testing aspects of performance not addressed in an objective format.

Make policy decisions regarding the information flow to candidates and other stakeholders. PES recommends that the Joint AICPA/NASBA CIC make decisions regarding the requirements for diagnostic feedback, the use of a pass/fail system versus scaled scores, and the use of composite scores versus subscores.

Establish a method for making a decision regarding the testing model to be implemented. Rather than endorsing the proposed CMT model, or suggesting an alternate model, PES believes that additional research must be conducted before a final determination is made regarding the
testing model to be implemented. Accordingly, PES developed a method to facilitate the eventual selection of a testing model. The proposed method includes considering up to 24 features of testing models when comparing the models and making a final selection. Various psychometric-, stakeholder-, and cost-related drivers that interact with the features are identified and should be considered in evaluating the models.

Implementation of this method may lead to the elimination of several models from any further consideration. Then, using the comparison features, the strengths and limitations of the remaining models should be evaluated more closely. Further research may be required to make refined determinations regarding the interactions among the features and the remaining models as an aid to final decision-making.
OBJECTIVES OF THE FEASIBILITY STUDY

Professional Examination Service (PES) conducted a feasibility study on behalf of the American Institute of Certified Public Accountants (AICPA) regarding the computerization and administration of the Uniform CPA Examination in 54 jurisdictions. The form and content of the proposed computerized examination are based on independent research and input from various constituencies, and the results are currently identified as the “ideal exam.” The proposed structure of the computer-based examination is a two-section examination. Each section would be taken and graded independently on a pass/fail basis. Testing for each section would be offered by appointment at regular intervals throughout the year in all jurisdictions. Candidates would be free to choose the order in which they take the sections. Diagnostic feedback would be provided to the candidates.

The main features of the ideal examination are:

- The first section would primarily test the basic knowledge needed by entry-level CPAs using multiple-choice questions. Day 1 is envisioned as a computer mastery test (CMT) wherein items are administered in groups known as testlets. Test length is variable, which means that some candidates are tested longer than other to determine if they pass or fail.

- The second section would emphasize assessing the skills needed by entry-level CPAs using performance assessment. Day 2 would consist of four simulations, about 1½ hours each in length, that would simulate real-world practice and assess higher-level, integrated knowledge and skills, including analytic, research, and communications skills.

The feasibility study addressed four sets of issues with regard to the ideal examination:

- Time frames for computerizing the ideal examination, including
  —the feasibility of computerizing and administering the Day 1 examination, using a computer mastery testing (CMT) model, by Spring 2003
  —the feasibility of computerizing and implementing the Day 2 examination, comprising four simulations including free-response questions, by Spring 2003

- Financial viability of computerizing the ideal examination, including
  —cost to develop an objectively scored item
  —cost to develop a testlet
  —cost to develop and computerize a simulation, including the cost of grading free-response questions
  —cost to computerize the ideal examination (including hardware/software)
  —cost to administer (deliver) Day 1 and Day 2 of the ideal examination

- Assessment of the technical viability of the computerized ideal examination, including
  —psychometric evaluation of the CMT model
  —psychometric evaluation of the simulation component of the ideal examination
• Assessment of stakeholders' reactions to the computerized ideal examination, including candidates', boards', and practitioners' acceptance

Over the course of the conduct of the feasibility study, PES identified significant technical and stakeholder concerns with regard to the feasibility of implementing the ideal examination. Accordingly, PES identified alternate testing models and key features of these testing models which might be employed to describe and compare the models and to make a final decision regarding the format of the computerized version of the Uniform CPA Examination.
METHODOLOGY FOR THE CONDUCT OF THE FEASIBILITY STUDY

Data Collection Initiatives

To conduct the feasibility study, data were gathered from subject-matter experts (SMEs) representing the AICPA, the National Association of State Boards of Accountancy (NASBA), and individual state boards; and from external sources including psychometricians, computer-based testing (CBT) vendors, test developers, test administrators, and other professionals familiar with issues relevant to the development and implementation of the AICPA’s ideal examination. PES reviewed research literature and AICPA, NASBA, and state board documents, and interviewed representatives of various credentialing agencies currently developing or implementing CBT programs.

Discussions with psychometricians and test developers focused on the following topics:

- Day 1 (CMT examination composed of objectively scored items): advantages and disadvantages of various testing models, e.g., CMT, linear, computer adaptive sequential testing (CAST), domain sampling, computer adaptive testing (CAT); length of testing session; procedures to develop and pilot-test items and assemble testlets and examinations; hardware and software requirements; potential for content coverage; reliability (and standard error of measurement); efficiency; size of item pool; generation of diagnostic reports; standard setting; candidates’ reactions and other stakeholders’ responses; security; cost.

- Day 2 (simulations): time needed to develop the program; procedures to develop the program; procedures to develop and pilot-test items and assemble simulations; general scoring issues (process versus content); potential for content coverage; relationship to objectively scored assessments; types of simulations (linear versus branched); level of interactivity; hardware and software requirements; standard setting; equating; generation of diagnostic reports; validity; candidates’ reactions and other stakeholders’ responses; security; cost.

See Appendix 1 for a list of the psychometricians and test developers interviewed as well as brief descriptions of their areas of expertise. See Appendix 2 for a list of exemplars of the research literature reviewed as part of the conduct of the feasibility study.

Discussions with CBT vendors focused on administering the ideal examination to large numbers of candidates, the availability of secure networks, and the availability of hardware and software to support both Day 1 and Day 2 of the ideal examination. See Appendix 3 for a list of the vendors who were interviewed or from whom resource costing documentation was obtained regarding the delivery of the computerized version of the Uniform CPA Examination.

Discussions with representatives of credentialing agencies focused on candidates’ reactions and other stakeholders’ responses to the implementation of a CBT program, including the effects on credentialing agencies and licensing boards. Discussions with representatives of information...
technology companies provided insight into CBT in the business and industry sector. See Appendix 4 for a list of the credentialing agencies and information technology companies whose representatives were interviewed or from whom resource information was obtained.

Finally, ongoing discussions with individuals connected to the AICPA, NASBA, the Joint AICPA/NASBA CIC, and the current examination program led to clarification of PES's understanding of the ideal examination.

Clarification of Associated Facts and Assumptions

To evaluate the feasibility of implementing the Uniform CPA Examination in a computerized environment by the year 2003, PES identified numerous facts and assumptions regarding the administration of the ideal examination, cost factors, and the roles and responsibilities of the boards of accountancy. The following narrative summarizes these facts and assumptions and highlights points wherein departures from the assumptions may have a significant impact on feasibility in terms of estimating resource requirements and costs. (Appendix 5 contains additional details regarding these facts and assumptions.)

General assumptions regarding the administration of Day 1 and Day 2 of the examination.
PES made assumptions regarding (a) total testing time per candidate on Day 1 and Day 2 of the ideal examination, (b) the sequence of events regarding candidate registration and testing, (c) the use of testing windows for the scheduling and administration of the computer-based tests, (d) the implementation of uniform requirements across jurisdictions regarding such policies as conditioning, (e) the number of candidates “in the system,” and (f) the number of candidate “retries” permitted per year and in total.

The impact of any inaccuracies regarding assumptions a through d is generally clear; the impact regarding resource requirements and costs can be readily estimated. Total testing time relates directly to costs for seat time; extending the duration of either testing day will increase the cost for the testing day. The sequence in which candidates register should not impact the feasibility of implementing the examination program. In regard to the use of testing windows and the imposition of uniform requirements across jurisdictions, violations of the assumptions will impact resource requirements and costs. For example, not implementing testing windows or uniform conditioning requirements across jurisdictions will increase the administrative complexity of the program and the requirements regarding quality controls.

The impact of inaccuracies in regard to the related assumptions e and f is more difficult to determine. That is, an accurate estimate of candidate demand for the test is critical to plan for the required CBT vendor build-out, to develop adequate numbers of test items and examination forms, and ultimately, to estimate the overall cost of the examinations and the cost to the candidate. In order to verify the accuracy of the assumptions made in regard to the number of candidates in the system, and the number of retries permitted per year and in total, the following additional information is required:
1. The absolute number of candidates sitting for the Uniform CPA Examination annually;
2. The average number of “total tries” per candidate;
3. The percentage of candidates who currently retest as frequently as permissible (i.e., twice each year); and
4. The percentage of candidates who drop out of the testing process before passing the entire examination.

PES understands that information which may inform these assumptions has been collected from Spring 1999 candidates sitting for the Uniform CPA Examination.

The following additional market research is strongly recommended in order to adequately estimate the absolute number of candidates who will be in the system annually, and the demand they will place on the system:

1. How much will candidates be willing to pay for the examination?
2. If given the opportunity, will candidates be willing to pay more and test more frequently than they currently do?
3. Can the examination sponsor create an increased value for the credential and thereby increase the candidate volume?

**General assumptions regarding the scoring of the Day 1 and Day 2 examinations.** PES made assumptions regarding (a) the requirement for separate pass/fail scores for Day 1 and Day 2 of the ideal examination, (b) the requirement for diagnostic feedback for failing candidates, (c) the development of policies regarding the release of scores subsequent to the day of test administration, (d) the development of a “banking” system whereby candidates might carry over examination scores for a set period of time, and (e) the Day 1, Day 2, and cumulative passing rates for first-time test takers, repeat candidates, and 4- and 5-year degree candidates.

On the basis of information gathered from key stakeholders, PES is confident that inaccuracies with regard to assumptions a through d will have minimal impact on the conclusions reached concerning the feasibility of developing and implementing the ideal examination.

Because the ideal examination is designed to be compensatory with regard to the scoring of Day 1 and Day 2, it is not possible to project the passing rates for either day, or for both days, for first-time test-takers and for repeat candidates without conducting empirical analyses of the test scores of current candidates on the four separate sections of the examination. Moreover, by the year 2003, many of the jurisdictions will require the candidates to complete a 5-year academic program. The shift in the level of preparation of the candidates further compounds the difficulty of accurately projecting passing rates of future candidates.

As a result, PES is not confident about the accuracy of the assumptions made about the passing rates on the ideal examination and the impact on the candidate volume. Moreover, inaccuracies in regard to these assumption may affect the conclusions reached regarding the feasibility of developing and implementing the ideal examination. Since the passing rates may be a lightning
Because of stakeholders' reactions and may directly impact candidate volumes, PES recommends that these assumptions be investigated.

The following additional information concerning first-time test takers, repeat candidates, and traditional and 150-hour degree candidates is required to refine the assumptions regarding passing rates:

1. The impact of implementing a fully compensatory pass/fail scoring system on Day 1 (based on candidate performance on the current multiple-choice questions (MCQs)) and other-objective-format items across the four separate sections of the examination.

2. The impact of implementing a fully compensatory pass/fail scoring system on Day 2 (based on candidate performance on the current essay questions across all sections wherein essays are administered).

3. The cumulative impact of applying Day 1 and Day 2 pass/fail decisions, e.g., How many first-time candidates and repeat candidates will pass both parts? How many will fail both parts?

Specific facts and assumptions regarding Day 1 of the ideal examination. PES understands that Day 1 of the ideal examination includes (a) the use of a variety of graphically enhanced, selected-response questions, (b) questions timed at 108 seconds each, (c) the use of a CMT model, including a 100-item base test and successive 25-item testlets, and (d) the application of stopping rules after the base test or successive testlets. PES has made assumptions regarding (e) the percentage of candidates for whom a pass/fail decision might be made after the base test and after each successive testlet.

On the basis of the information we have gathered from the AICPA professional staff and other key stakeholders, PES is generally confident that the financial impact of variations in facts a–d can be estimated. For example, if candidates should require either more or less time per question, the costs related to seat time might increase or decrease, respectively.

On the other hand, PES recommends further investigation of all assumptions concerning the percentage of candidates for whom a decision may be made after the base test and each successive testlet, as the rate of item exposure directly affects all cost and resource requirement estimates. Accordingly, the following additional information is required to develop accurate cost and resource estimates:

1. Empirical estimates of the actual percentage of candidates for whom a decision might be made after the administration of the base test in the morning session.

2. Empirical estimates of the actual percentage of candidates for whom a decision might be made after the administration of each additional testlet.
Specific assumptions regarding Day 2 of the ideal examination. PES made specific assumptions regarding (a) the sequential administration of the four 90-minute simulations, (b) the development of a 30-minute tutorial, (c) the development of practice materials to be made available via CD-ROM, (d) the use of a validated test blueprint to guide examination construction initiatives, (e) the incorporation of a text- and graphics-based scenario within each simulation, (f) the requirements for the use of on-line software such as spreadsheets and databases, (g) the requirements to score communications and writing skills as part of the simulations, and (h) the use of live and automated readers to score the written responses of the candidates.

To calculate the cost of developing and implementing the simulation-based assessments, numerous assumptions have to be verified regarding the costs required to support the research and development process. For example, will students and/or newly certified CPAs be available for purposes of pilot-testing during the development phase? If the new test specifications support the assessment of communications skills, including writing, to what extent will the scoring of free responses be integrated into the assessment? Will the responses be scored for both form and content?

Without directed research, it is not possible to develop final cost and resource estimates for the development and implementation of Day 2 of the ideal examination. Accordingly, the estimates of cost and resource requirements PES has provided must be considered preliminary.

Specific assumptions regarding the development and administration of a linear examination. To provide a benchmark for evaluating the cost and resource requirements for Day 1 of the ideal examination, PES made assumptions regarding the development of a linear examination. These assumptions included (a) the variety of graphically enhanced, selected-response questions to be included in the linear examination, (b) the testing time per question, (c) the number of test forms to be published per year, and (d) the total testing time per candidate.

All estimates regarding cost and resource requirements will be impacted to the degree that more or fewer test forms are published per year, and total testing time is determined to be more than or less than 7 hours.

Assumptions regarding costing. PES made assumptions regarding (a) the cost components associated with development versus the cost components associated with ongoing administration, (b) the use of ranges to estimate the simulation development costs, (c) the inclusion of software costs for individual boards, (d) the exclusion of capital expenses for testing networks, (e) the use of a single test network for delivery, (f) costs for expert-graded essays, (e) per item development costs for selected-response questions, and (g) computer-based administration costs.

Separate budget schedules were prepared reflecting development costs and ongoing administration costs for Day 1 and Day 2 of the ideal examination. Due to the uncertainties identified in regard to the development of the simulations, PES used cost ranges rather than estimated costs; these cost ranges were established on the basis of discussions with sponsors of at
least four other credentialing programs that are either developing or implementing simulation-based examinations.

While the costs of developing and implementing the examination in a computerized environment can be estimated, additional financial issues impact the ultimate price of the ideal examination to the candidates. For example, costs to be determined by third parties responsible for various aspects of administration and delivery will impact the overall cost of the examination. Moreover, boards of accountancy make final decisions on pricing.

**Overall assumptions concerning the roles and responsibilities of the boards of accountancy.** PES has made assumptions regarding the process by which candidates will proceed through the credentialing system, including assumptions regarding (a) the continued determination of candidate eligibility and final licensure by jurisdictions, (b) the use of a centralized database and specially designed software to enter, transmit, and receive candidate data, (c) the use of uniform conditioning requirements across the jurisdictions, (d) the use of unique candidate identification numbers, (e) the flow of information between the candidates and the jurisdictions, and between the candidates and the CBT vendor, (f) the flow of information between the AICPA (or other designated agencies) and both the CBT vendor and the essay-grading vendor(s), and (g) the electronic transmission of all score data and results.

PES perceives that the assumptions made in regard to the boards of accountancy are reasonable; however, while issues such as conditioning have been addressed, no final answers have been developed. Inaccuracies in the assumptions would have varying impact on the cost and resource requirements underlying the development and implementation of the ideal examination; for example, should the states not adopt uniform conditioning requirements across the jurisdictions, the overall cost and resource requirements would increase.

---

2 In addition, any decision by stakeholders to use more than one network (e.g., one operated by a current CBT vendor, a jurisdictional-based network, or an education-based network) will increase cost and resource requirements in regard to such things as quality control initiatives, communications between networks, and security concerns.
EVALUATION OF THE IDEAL EXAMINATION

PES has evaluated the ideal examination in terms of technical issues and stakeholder perspectives, as well as the feasibility of developing and implementing the examination in terms of cost and timing. The following sections document the results of that evaluation for Day 1 and Day 2, and for the implementation of the comprehensive 2-day examination.

Day 1: Computer Mastery Test

CMT seeks to determine the mastery status of a candidate rather than estimate a score. After some minimal-length test (the stage 1 or base test), the computer makes one of three decisions incorporating the relative seriousness of two possible decision errors. The decision errors are referred to as decision error A (false positives), passing a candidate who should fail, and decision error B (false negatives), failing a candidate who should pass. The three possible decisions are pass the candidate, fail the candidate, or continue to test in order to create the opportunity to pass or fail the candidate with minimal risk of decision errors A or B.

For candidates who are determined to be clear passers after the base test, there is little risk of decision error A, and for candidates who are deemed to be clear non-passers after the base test, there is little risk of decision error B. Accordingly, testing is stopped. For the candidates for whom base test performance does not allow decisions to be made (without substantial risk of either decision error), testing continues. The computer administers brief, randomly selected examinations, known as testlets, until a specified level of confidence in the decision outcome is reached. Each testlet consists of items of comparable difficulty and content to those of the base test. Testing continues until a decision can be made or until the last of the testlets is completed, whichever comes first.

In a typical administration, there is a scoring range in which testing continues beyond the base test. Candidates whose scores are either above or below the scoring range pass or fail, respectively. A decision is made after each testlet to classify the candidate or to administer another testlet if the candidate's performance remains indeterminate and the number of completed testlets is fewer than the maximum number available.

A central concern in CMT is that of maximizing the probability of making a correct decision while minimizing test length. The testing model provides for shorter tests for candidates who have mastered (or not mastered) the content, and longer tests only for candidates whose status is indeterminate. Accordingly, exposure of items within the pool may be minimized, reducing the number of items which need to be developed. Other high-stakes credentialing examination programs have indicated that a pass/fail decision may be reached after the administration of the base test for more than 50% of the candidates.

A feature of CMT is that the base test and the testlets are assembled prior the actual computerized administration. Control over problems such as item-ordering and context effects (i.e., unintended cueing) is maintained through review of the pre-assembled forms. Both the base
test and the testlets are constructed to be content-balanced, with a spread of difficulty.

**Technical Perspective on Day 1, CMT**

In evaluating the use of a CMT model in a high-stakes credentialing environment, two related issues require a resolution to enhance the viability of the Day 1 assessment for “morning only” candidates: the provision of diagnostic feedback to candidates who fail on the basis of performance in the morning session, and the adequacy of content coverage for those candidates. These issues are described in regard to Day 1, followed by recommendations to enhance the psychometric characteristics of the assessments. Finally, in accordance with the specific CMT model proposed for the Day 1 assessment, additional specific psychometric issues are identified and described. Recommendations addressing the perceived deficiencies are offered, where available.

**Diagnostic subscores for “morning only” candidates.** On the basis of two general assumptions of the CMT model underlying the Day 1 test, candidates whose morning session scores fall below the indeterminate score range cease testing after the base test. Accordingly, “morning only” candidates will be administered 100 items. Depending on the complexity of the final test specifications, it may be unlikely that 100 items will be sufficient to yield reliable diagnostic subscores in all areas of the specifications.

One means of addressing diagnostic feedback for failing “morning only” candidates is to continue testing beyond the point of making a dependable pass/fail decision until sufficient information is collected to develop effective diagnostic score reporting. The additional items could be selected to comprise a diagnostic test assembled specifically for individuals who fail the base test assessment. Because Day 1 candidates must anticipate the possibility of testing for the entire day, scheduling “morning only” failing candidates for a brief afternoon session would not be administratively difficult, although it would require the development of additional testlets organized by content.

**Content coverage for “morning only” candidates.** A second psychometric issue involves the adequacy of content coverage for the Day 1 examination. Candidates testing during only the morning session on Day 1 will respond to 100 operational items. The content validity of the Day 1 assessment may be diminished for “morning only” candidates, because the number of test items reflecting certain content areas within the test specifications will be minimal. Test length needs to be long enough to make the case for content validity.

Two potential solutions are recommended to address this issue for the Day 1 examination. One approach would involve increasing the minimum number of items administered to candidates before a pass/fail decision is reached. Expanding the length of the Day 1 base test would improve its content coverage and content validity. (One high-stakes licensure examination has recently undergone conversion from a 2-day examination to a 180-item linear examination.)
A second approach would require the administration of additional items in content areas not adequately tested before a pass/fail decision is made. If this approach were pursued, automated test assembly techniques could be used to build testlets that emphasize particular content dimensions and psychometric properties (Lewis & Sheehan, 1990; Wainer & Kiely, 1987).

**Content coverage.** While the 100-item base test may permit adequate content coverage as outlined in the test specifications, each subsequent 25-item testlet is definitely too short to include all of the details of the test specifications. Accordingly, different testlets may not be truly parallel, as theoretically suggested by the CMT model, further confounding the problems associated with the provision of diagnostic feedback to failing candidates. Content coverage for afternoon test takers may be adversely impacted by random omissions in successive testlets.

**Item exposure and test security.** Each successive stage of Day 1 will expose items to a smaller number of candidates; however, the items allocated to the morning session will be administered to all candidates. The validity of the Day 1 assessment may be compromised by security threats due to the overexposure of items administered during the early stages of the test.

To correct this problem, the administration of items within testlets should be actively monitored with the goal of exposing each item to an equal number of candidates. The testlets should be rotated among morning and afternoon testing sessions to achieve this goal, in conformance to item exposure rules to be established.

Pretest items will be especially subject to overexposure, if they are restricted to the morning session. Careful monitoring of the number of candidates who have viewed pretest items and item rotation strategies must be used to avoid compromising the validity of Day 1 assessment items.

**Testlet and item discrimination.** To be effective, the items and testlets created must yield maximum discrimination at the ability level near the passing score. The nature of the CMT model will place a significant demand on item writers to generate highly discriminating items at a very specific point on the test score scale.

The use of item models for item writers may promote the development of highly discriminating items. If necessary, a smaller number of highly discriminating testlets could be assembled, but their content is less likely to be parallel. If this strategy is employed, it would be necessary to evaluate the impact of a smaller number of testlets on item exposure controls and testing frequency.

**Indeterminacy of passing status.** Unlike linear testing, CMT attempts to make a decision regarding mastery status by reducing the range of a candidate's ability estimate until it is clearly above the passing score or clearly below it (Folk & Smith, 1998). The model proposed for the Day 1 assessment assumes that 250 operational items (i.e., the base test in the morning plus six testlets in the afternoon) will be sufficient to yield a determinate score and an accurate pass/fail decision for the majority of candidates. If the item bank is unable to support the construction of testlets that adequately discriminate among candidate ability levels, a large number of candidates
may have indeterminate scores and thus require reassessment. This outcome would further strain the item bank because of item exposure control parameters.

Simulations should be conducted to assess the relative probability that a large number of candidates will be indeterminate at the conclusion of Day 1 testing. It should be possible to obtain an approximate estimate of this outcome with the current CPA item banks. If the outcomes of these studies suggest that a large number of Day 1 candidates will have indeterminate pass/fail decisions at the maximum test length, the length of the Day 1 assessment could be increased to minimize the probability of an indeterminate decision. A different strategy to resolve this problem would involve administering a more discriminating testlet to individuals whose initial test performance would suggest an indeterminate pass/fail decision. (Luecht and Nungester (in press) have described this testing strategy as Computer-Adaptive Sequential Testing [CAST].) This strategy has not been implemented in any operational credentialing programs, to date.

**Standard setting.** There is an extended research literature regarding standard setting methods in connection with CMT. One prevalent method is based on a variation of procedures recommended by Angoff (1971). This method would require a panel of content specialists to serve as judges for the standard setting study. After an orientation and training period, judges would review each item on one form of a complete, 250-item, Day 1 test and then estimate the probability that a minimally competent CPA candidate would answer the item correctly. Judges would supply these estimates for each item on the 250-item test. To compute a passing score, the judges’ item ratings would be summed among items, and then an average passing point for the total test would be computed by dividing the sum by the total number of judges.

This raw passing score would be expressed on a scale, called the theta scale, reflecting candidate ability level. Statistical data collected during pre-implementation calibration studies would permit this conversion. The purpose of this score conversion would be to derive a passing score, expressed on a common ability scale (i.e., the theta scale), that could be applied to each form of the CMT.

A candidate’s pass/fail status would be determined on the basis of his or her position on the theta scale after the administration of the morning base test, and, if necessary, after each successive testlet during the afternoon testing session. A level of confidence, based on the standard error of estimate associated with a candidate’s score after each testing period, would be established to help guide pass/fail decisions. For example, testing might cease if a candidate’s theta estimate were above (or below) the passing score and the amount of error associated with the candidate’s score estimate did not include the passing score. Calculation of the standard error of estimate of a candidate’s score and a comparison of the candidate’s theta estimates to the passing score would be made after each testing segment was completed, and testing would cease if the estimate was sufficiently accurate, the last testlet was completed, or testing time had expired for the last testlet.
If a candidate completes all testlets and the level of precision does not meet the stopping rules in effect for the Day 1 assessment, a pass/fail decision would be reported, based on the candidate’s final theta estimate relative to the passing score. It is recommended that the test stopping rules, including the amount of precision associated with candidate score estimates, be determined on the basis of simulation studies conducted prior to the implementation of operational testing.

*Item calibration.* It is recommended that items be calibrated using Item Response Theory (IRT) techniques. Model-fit studies should be conducted to determine which IRT model best fits the item data. The goal of these studies will be to compile a pre-equated item bank that will support the CMT model and standard setting approach recommended for the Day 1 assessment.

**Stakeholder Perspective on Day 1, CMT**

*Use of a CMT model and variable-length testing sessions.* Several significant stakeholder concerns have been identified with regard to CMT. It has been noted by a representative of one national licensure examination program which had previously implemented CMT, that the general public does not understand the testlet development process or CMT’s IRT underpinnings. Accordingly, candidates, boards, and practitioners do not fully appreciate that at a certain point (short of a candidate’s completing the entire test), the likelihood of that candidate’s passing the test becomes so small that it makes little empirical sense to continue testing.

PES has identified this perceived fairness issue as of special relevance to the Joint AICPA/NASBA CIC. While it is not possible to estimate the absolute number of candidates who will fail after the morning base test, the number is likely to be as high as 25% of the total candidate population taking the Day 1 assessment. Accordingly, many candidates will have the experience of having been administered a “very short test” before being terminated. For these candidates, the shift from a 2-day examination to a 3-hour examination with comprehensive content coverage may seem to be very unfair and very arbitrary.

PES anticipates that there may be concern on the part of boards of accountancy and the practitioners with regard to the reduction in testing time associated with the ideal examination. That is, the total time to test may be less than the current period, even for those candidates requiring the maximum number of testlets. For the candidates who pass Day 1 on the basis of the morning session or on the basis of the morning session and part of the afternoon session, total testing time would be considerably reduced. Accordingly, PES recommends that the AICPA conduct and disseminate the results of pilot-test studies which evaluate the content coverage of the proposed examination in comparison to the content coverage of the current examination.

*Content coverage.* The use of a compensatory scoring model for Day 1 suggests that key stakeholders may question the utility of that type of examination to identify candidate strengths and weaknesses in relation to content, in comparison to the use of the four sections of the examination as currently structured. The conduct of the full-scope practice analysis (to be undertaken in 1999 – 2000) should include attention to those elements of the current examination which will be emphasized or de-emphasized in the new test specifications underlying the ideal
examination. In addition, questions of dimensionality within and across the four sections of the current examination must be addressed in order to support the use of a single score and in response to stakeholder concerns about the perceived value of the new examination. PES is aware that studies of the dimensionality of the Uniform CPA Examination are under way; those results should inform policy decisions in regard to the use of a single compensatory score for each of the 2 days of the ideal examination.

**Timing: The Feasibility of Implementing Day 1 by 2003**

The results of PES's research indicate that Day 1 of the ideal examination can be implemented by 2003. At least one testing agency has proprietary software drivers for CMT test assembly, and other CBT-delivery vendors have stated that they would develop such drivers for the AICPA.

On the other hand, the demand for "seat time" may be beyond that currently available. PES is aware of major CBT build-outs occurring. In order to guarantee seat time, the AICPA must contractually obligate the CBT vendor to provide adequate access to candidates.

PES is aware that the AICPA is in process of conducting a targeted practice analysis study in regard to general business knowledge, in particular, and a full-scope practice analysis study in regard to the practice of the profession, in general. It will be very important to have the results of these studies available, so as to ensure that all item development efforts are targeted to contemporary practice.

PES also knows that the AICPA is in process of implementing enhanced test development initiatives so as to increase the supply of objectively scored items. The absolute number of objectively scored items required for CBT is far greater than the number required for the paper-and-pencil administration of the examination. PES recommends that the AICPA continue its process of "ramping up" test development efforts in order to produce sufficient quantities of items for use in a CBT environment.

Finally, PES strongly recommends the conduct of research studies wherein estimates of pass/fail rates related to the compensatory nature of Day 1, and pass/fail rates related to the use of the CMT model might be determined. The results of such studies are necessary to refine the estimates of the numbers of items and/or test forms to be developed for the ongoing implementation of the Day 1 program, regardless of the testing model.

**Day 2: Simulation Test**

Simulations and other performance assessments demonstrate excellent face validity, meaning that they appear to be related to the skills needed to perform well in a given occupation or profession. Many important aspects of these assessments must still be evaluated, however, to determine if the scores yielded are, in fact, a valid reflection of the knowledge, skills, and abilities that the assessments are intended to measure.
Content coverage. The validity of any examination may be questioned if candidate’s performance is significantly affected by factors unrelated to the purpose of the examination or if key aspects of performance are not represented. Although both of these features apply to performance assessments, in general, the latter is of particular significance to the simulations, since the time to complete each assessment in the ideal examination is relatively long, limiting the number of simulation problems that can be administered to only 4. Unlike multiple-choice tests that use many items and hence permit good content coverage, the number of performance assessments may be insufficient to cover all of the important aspects of the skills to be measured (Messick, 1994).

Further, in regard to content coverage, Crocker (1997) has stated:

Studies of content representativeness of performance exercises should provide empirical evidence that such exercises and their scoring rubrics represent a faithful translation of the objective or domain specifications into tasks and scoring criteria. The justification for using such assessments for any type of high stakes decisions requires evidence that the assessment content (both exercises and rubrics) represents more than an idiosyncratic interpretation of the domain specifications . . . Thus, content judgments of performance assessments are more complex to plan than those of assessments in more traditional formats. Additional considerations . . . include:

1. What structural features of the performance exercise should be evaluated?
2. What additional precautions for test security are needed?
3. What additional criteria should expert judges use in evaluating exercise content?
4. How are scoring rubrics incorporated into the review?
5. How can reliability (generalizability) of experts’ ratings be estimated?

Accordingly, it is recommended that the number of simulations be increased and the level of complexity be reduced in order to address the concern regarding the limitation of content coverage in the Day 2 assessment. Expanding the number of simulations will facilitate the development of truly independent measures, thus improving the reliability of the assessment.

The development of less complex simulations will also enhance the breadth of content coverage by allowing a larger number of skills to be assessed in multiple contexts.

Content validation studies. The complexity of the simulations, their format, and the use of the on-line supplementary materials (e.g., authoritative research literature, spread sheet programs), the restricted number of exercises, the enhanced need for security for the highly memorable content, and the complex scoring rubrics are all features that complicate the conduct and design of content judgment studies. However, the implicit assumption that the simulations will represent an authentic approach to adequate and appropriate content coverage requires
empirical validation. Accordingly, PES recommends that appropriate content validation studies be incorporated into the development phase for Day 2.

Scoring. Another factor in the design of problems concerns the development of scoring rubrics. Whatever the scheme that is adopted for scoring the simulations, the first step will be to identify the elements in the candidate performance that are to be considered in the score. Decisions will need to be made about how much of the performance to capture in a computer record and which aspects of performance to include in an evaluation of the adequacy of the responses. These decisions are ultimately subjective and may reflect the biases of the individuals making these judgments (Crocker, 1997).

Scoring of communications skills, including writing. Currently, the AICPA uses experts to read and grade the writing tasks required in the Uniform CPA Examination. As described, Day 2 of the ideal examination also includes essays in the performance evaluation of candidates. Numerous technology-based solutions are being developed (or have recently come on-line) which may provide a partial solution to the scoring of such tasks. Bennett (1998) notes that research in the 1990s focused on the use of automated scoring of constructed responses in connection with mathematics, computer programming, architectural design, medical problem solving, and writing. Now, they consider automated essay grading the innovation that has the broadest applicability to high-stakes testing programs. While reports of such solutions may be overly optimistic, they indicate that automated scoring holds great promise and should be evaluated for its potential use in the grading of communications skills, including writing.

Currently, at least five automated scoring programs for essays are available (Bennett, 1998):

1. Project Essay Grade (PEG) (see Page, 1967; Page & Petersen, 1995)
2. Intellimetric Engineer (1997)

As Bennett has said, “data are available on PEG, Intellimetric, E-rater, and the Intelligent Essay Assessor. Without exception, these data indicate that automated essay scoring programs agree as highly with human raters in judging essays as human raters agree among themselves.” (See also Burstein, Kukich, Wolff, & Lu, 1998; Elliot, Burnham, Chernoff, & Kern, 1997; Landauer, Laham, Rehder, & Schreiner, 1997; Page & Petersen, 1997.)

It is not clear, however, that stakeholders, including candidates, boards, and practitioners will accept the use of automated essay scoring in high-stakes credentialing programs. Bennett & Bejar suggest that one solution might be that the high-stakes credentialing programs replace one of the two human readers with a machine score, while arbitrating discrepancies between machine and human scores with a second human grader. (This solution is consistent with the AICPA’s desire to continue the use of human grading, albeit at a reduced level.) A second solution would
be to use two different automated graders (each employing a different analytic technique). If the two automated graders agree on the score, the score might be accepted, and if the two automated graders disagree on the score, a human rater might arbitrate the score. A third solution to the potential discomfort of the stakeholders to reliance on automated scoring might be to have the automating graders perform an initial "cut" on the essays. While the responses of clear "passers" or "failers" would not need to be reviewed, their papers might be randomly audited. On the other hand, the responses of marginal candidates might be routinely rescored by human graders.

In any case, PES recommends that the AICPA continue to explore the use of automated grading. Before a final determination is made as to the specific solution, the AICPA should undertake the conduct of limited pilot-testing to evaluate the reliability of automated scoring in connection with bullet point responses, short essays comprising 3 to 5 concepts, and long essays comprising 6 to 10 concepts. At the same time, the pilot-testing should be designed to evaluate the utility of the variety of available programs.

**Combining scores from different vendors.** Should the ideal examination include writing tasks, detailed quality control mechanisms will be required to transmit data, including candidate responses and scores, between the CBT vendor, the automated grading vendor, and the agency responsible for determining the candidates' total scores on the Day 2 examination.

**Reliability.** Reliability of measurement is also likely to be low for performance assessments, especially if the number of simulations is limited. Accordingly, the performance of a candidate on one simulation may not correspond well with his or her performance on another. This may be attributable to subjective scoring procedures, the small number of the behaviors sampled in the simulations, or a performance too specific in regard to the situation in which generalization is desired (Dunbar, Koretz, & Hoover, 1991; Linn, Baker, & Dunbar, 1991).

Consequently, many different problems may be needed to form a good evaluation of the candidates' abilities. The National Board of Medical Examiners found that approximately 10 simulations were required to yield adequate reliabilities. Others report using 15 – 20 simulations to ensure reliability. The National Council of Architectural Registration Boards replaced one 12-hour simulation with 15 simulations administered in three different examinations, each of which must be passed.

PES recommends that the format of the simulations be modified so that the number of simulations can be increased to improve the content coverage (validity) and the reliability of the derived score data.

**Standard setting.** At least three different methods for setting standards with complex performance assessments are discussed in the literature. These three methods are the iterative judgmental policy-capturing method (Jaeger, 1995), the dominant profile judgment method (Plake, Hambleton, & Jaeger, 1997), and the contrasting groups method (Clauser & Clyman, 1994). These procedures are generally all methods for modeling the decision making process of human judges. PES recommends that the AICPA conduct targeted research regarding their applicability to the simulations as described in the ideal examination.
Equating. Although this question has been the subject of much research, the equating of performance assessments remains a largely unresolved issue. Despite a general lack of satisfaction with the methods proposed, the use of performance assessments in large-scale testing programs, primarily in schools, has required that some method of equating be performed. The methods that have most often been used are (a) those based on total test scores and using classical procedures and (b) those using IRT, calibrating the performance assessments using a partial credit model. Overall, the equating of long simulations may be difficult given the multidimensionality of the assessments, that is, the inclusion of writing skills, research skills, computational skills, etc.

Classical procedures: In the instance where multiple-choice items are administered with the performance assessments, the multiple-choice items can serve as common items to equate total scores made up of both multiple-choice and performance assessments using traditional methods. If the score is to include only performance assessments, the multiple-choice items can be used to adjust for differences in the difficulty of the different performance assessments as well as differences in the abilities of the candidate samples taking the different test forms. These methods assume that the multiple-choice items are measuring the same constructs as the performance assessments. If only performance assessments are administered, equipercentile methods can be used after smoothing the total scores. An assumption must be made here that the groups taking the two examinations are equivalent in ability.

IRT methods: In IRT, tests can be equated through their item statistics, again with linkages between forms established with previously calibrated items. If both multiple-choice and performance assessments are administered, they can be placed on the same scale using partial credit models to calibrate the performance items. Later forms can be placed on the same scale through use of the previously used multiple-choice items. If no multiple-choice items are administered, the new form must contain performance items that were previously calibrated, in order to get the new form onto the scale. This requirement may be difficult to meet if the new form contains few performance assessment items.

Diagnostic feedback. The Day 2 test will consist of 4 lengthy simulations covering various aspects of the test specifications. Most lengthy simulations are interrelated to the degree that scales derived from sections of a simulation are interdependent. Interdependent scales prohibit the reporting of separate scale subscores. Thus, the small number of independent scores available from the Day 2 assessment may not permit the reporting of diagnostic feedback for failing candidates.

One means of addressing the issue of diagnostic feedback for failing candidates is to increase the number of simulations. An increase in the number of independent simulation scores (based on process or content), would permit the generating and reporting of meaningful diagnostic feedback.
Pretesting. The use of shorter simulations simplifies the process of pretesting. For example, if 15 short simulations are administered to candidates, an additional simulation that is not scored could be included for purposes of pretesting under realistic conditions. A long simulation would be difficult to include as part of an operational examination program. This would mean that some other method of pretesting would be needed, including the identification of suitable examinees with some means of motivating them to give a good effort.

Fidelity. One of the important features of simulations and one of their advantages is fidelity to professional practice. Fidelity is never perfect, however. Performance on a simulation differs from performance in actual practice because the activities are carried out in a testing situation. One candidate may be nervous and overlook important features of the simulation while another is more careful than usual because the performance is being monitored. Moreover, a high degree of fidelity may not enhance the measurement properties of the simulations. Practical considerations suggest that while some degree of fidelity is important, beyond a point, there are rapidly diminishing returns.

Those with actual experience in developing simulations who were interviewed for this project all indicated that the temptation to include more fidelity than needed for the measurement should be resisted. A desire for high fidelity can lead to the development of problems that are too elaborate, with too many features. Realism may need to be sacrificed for practicality so that problems are manageable and scoring straightforward. In general, the more structured the problem, the better. Striving for high fidelity can also lead to simulations that include concepts that make the situations more realistic but are not considered important to measure.

General impact of using a limited number of simulations. Although reliability is an important technical issue, other considerations suggest that presenting more and shorter simulations may be a better use of time than presenting fewer and longer ones. In general, flaws in the simulations that inevitably arise during problem development become magnified because of the relatively great weight each simulation has on the total score. The use of more simulations lessens the impact of any one of them on the overall performance of a candidate.

Burden on the candidates. Complex simulations that permit different kinds of responses from candidates offer many opportunities for flaws in functioning to arise. Directions to candidates must be very clear so that errors do not arise out of simple misunderstandings. If careless errors are made by knowledgeable candidates, they may be misdirected throughout the simulation or they may realize their errors and spend too much time trying to recover. In complex simulations where parts are interrelated, a simple mistake could affect a candidate’s performance in a major way.

Security. Examination security will be difficult to maintain in regard to the Day 2 assessment because of the highly memorable nature of long simulations. PES recommends that the AICPA undertake systematic research in order to investigate the degree to which “cloned” simulations can provide candidates with novel stimuli.
Stakeholder Perspective on Day 2, Simulations

Response to Briefing Paper No. 1. In October 1998, the AICPA disseminated a briefing paper, *Conversion of the Uniform CPA Examination to a Computer-Based Examination*, to key stakeholders, including members and administrators of boards of accountancy in all 54 jurisdictions, educators, CPAs in public practice and business and industry, and state CPA societies. The briefing paper provided an overview of the computerization initiative; a description of the two-section examination and the content focus; highlights of such key features as pass/fail grading, grade reporting, frequency of administration, and conditioning; and a summary of evidence supporting the change to a computerized format and the change to a two-section examination. Test specifications for a sample testlet and two simulations were included as well. Stakeholders were invited to complete and return a questionnaire that invited comments on the proposed model, as well as on a variety of other issues related to the conversion to the computer-based examination.

Seventy responses to the Briefing Paper questionnaire were received. Of these 70, 11 responses represented the consensus of boards of accountancy; 19 represented responses from individual board of accountancy members and administrators; and 40 represented other interested parties, including CPAs in public practice, industry, and education.

The responses indicated general consensus that the project is feasible and that computerization of the Uniform CPA Examination should move forward. However, there were concerns that the design of the ideal examination may not be optimal and that the testing model needs to be further analyzed, empirically tested, and modified, if appropriate.

While about two thirds of the respondents rated the proposed model as very good or the best model, only about one half of the boards of accountancy and their individual members and administrators rated the model as very good. Three boards felt they could not respond to the questionnaire until they had additional information. Overall the boards were in favor of conversion to a computer-based examination (Joint AICPA/NASBA Computerization Implementation Committee, 1999).

An analysis of the open-ended comments indicated some confusion and/or inaccuracies in the respondents' overall understandings of the proposed examination, most notably with regard to the availability of diagnostic information, the potential for testing communications and other skills, and the difficulty level of the examination. Accordingly, the ratings and comments received from these stakeholders should be interpreted cautiously. Future briefing papers should be disseminated with enhanced descriptions of the proposed CBT initiatives.

The essay. It remains apparent that stakeholder response to the inclusion of a writing task within the Uniform CPA Examination is controversial. Several jurisdictions have gone on record as indicating that they do not support the use of essay-type questions. PES has reviewed the AICPA's analyses of the impact of the essay: the essay tasks, per se, do not greatly impact candidate performance. On the other hand, PES suggests that the inclusion of the essay requirement has become a focus for dissension. Accordingly, PES recommends that should the
writing tasks be included within the requirements of the ideal examination, the AICPA develop procedures to validate the specific skills requirements that each writing task places on the candidate. Does the task exceed that generally required of newly licensed CPAs? Is a written response the only form in which the communication might take place? Further, the agency responsible for scoring the Day 2 examination must routinely assess the impact of the inclusion of the writing tasks to provide assurance that those tasks are not overrepresented within the test specifications.

Timing: The Feasibility of Implementing Day 2 by 2003

PES's research indicates that the feasibility of implementing Day 2 by 2003 is somewhat problematic. The outcomes of the AICPA's pilot project regarding the efficient development and scoring of valid and reliable simulations are required to complete this evaluation. The highly memorable nature of simulations suggests that a procedure to author multiple variations of problems will have to be developed and validated, issues with regard to content coverage and equating of performance across simulations will have to be resolved, the utility of automated scoring procedures must be empirically investigated, and the hardware and software requirements for the delivery network must be identified and operationalized.

Implementation of a Comprehensive Two-Day Examination

The ideal examination represents a major departure from the form and content of the current examination structure. The proposed structure of the ideal examination is a two-section examination. The first section, Day 1 (fundamentals of knowledge), would primarily test the basic knowledge needed by entry-level CPAs; the second section, Day 2 (performance assessment), would emphasize the skills needed by entry-level CPAs using computer simulations. The two sections of the ideal examination are not hierarchical. Grading of the sections is independent, and candidates can pass one section or the other. The ideal examination includes the use of a pass/fail grading methodology for each day; candidates will not receive numerical grades. However, candidates would receive notification of their pass/fail status on each section of the examination. Passing of both sections is required for licensure.

PES has identified issues regarding implementation of the comprehensive examination which are not tied to either Day 1 or Day 2, but are a function of the new form and content of the examination.

The Conduct of a Practice Analysis

Within the past 18 months, the AICPA has implemented practice analysis studies to pinpoint content specification in three key areas: information technology (IT), general business knowledge (GBK), and integrated knowledge and cognitive skills. The IT Practice Analysis (completed in 1999) identified IT skills and competencies that CPAs need to perform in auditing and other attestation engagements. These changes will be reflected in the November 1999 Uniform CPA Examination. The GBK Practice Analysis (currently in progress and scheduled to be completed
in Fall 1999) is designed to identify essential knowledge and skills in such content areas as economics and business organizations, management accounting, and working capital. Changes identified through that study may be implemented in the November 2000 Uniform CPA Examination. The Integrated Knowledge and Cognitive Skills Practice Analysis (to be implemented in Spring 1999 and completed in Spring 2000) is designed to identify the analytical, communications, and research skills that are both necessary for the protection of the public interest and feasible to test in either a paper-and-pencil or CBT environment.

The results of each of these practice analysis studies, most especially of the Integrated Knowledge and Skills Practice Analysis must be incorporated into revised test specifications for both Day 1 and Day 2 of the ideal examination. The use of a CBT delivery system should provide far greater opportunity for testing the integrated knowledge and higher-level cognitive skills than would a paper-and-pencil delivery system.

**Evaluation of Content Coverage**

The content coverage for both Day 1 and Day 2 is to be comprehensive; that is, the required knowledge will be assessed on one day, and the required skills will be assessed on the other day. Since candidates will be able to register separately for each of the two days, different candidates will be exposed to different combinations of Day 1 and Day 2 test material. To reduce unintended overlap and/or gaps in content coverage, the test specifications underlying both days must be carefully coordinated. At the same time, the test items for both days must be reliably classified so as to facilitate the monitoring of examination content for fidelity to the test specifications. This coordination between sections represents a new dimension in regard to the AICPA’s current test development initiatives, as PES understands that the contents of the current sections of the examination are independent of each other.

PES recommends that the AICPA monitor candidate performance on all Day 1 and Day 2 items which test similar or related content. Because of the high resource and cost requirements associated with the simulations, it is more efficient to identify any content which can be reliably assessed in an objective format, and to test it in that format only.

**Impact of Compensatory Scoring on an Examination Comprising Two Sections**

While it is the case that candidates are to receive separate scores for Day 1 and Day 2, a passing score on each day is required for certification. PES strongly recommends that the agency responsible for scoring the examinations develop procedures regarding the systematic reporting of the results of candidates’ performances across both days. It is PES’s experience that despite the fact that modifications of existing programs may be based on valid and reliable procedures, when a credentialing agency modifies its programs (e.g., changes the content of the test, the length of the test, the standard), the changes may lead to heightened security.
Educational Materials Development

Above and beyond the development and implementation of a CBT program, credentialing agencies are required to educate the candidates to the form and content of any new examination program. In the case of the ideal examination, there is a clear need to develop written and computer-based instructional materials which will permit candidates to familiarize themselves with both the appearance of the objectively scored Day 1 items, and the form and appearance of the Day 2 simulations, including all performance requirements.

PES has had discussions with professionals responsible for the computerization and implementation of the program of the National Council of Architectural Registration Boards (NCARB), including simulations. In that case, the vendor responsible for the development of the simulations and the associated scoring system was responsible for the development of a CD-ROM package available to candidates, providing practice problems. In the case of the National Board of Medical Examiners (NBME), exemplars of the simulation problems were widely disseminated through available channels, including medical schools, so as to familiarize students with the experience of working with interactive branching simulation problems. Finally, in the case of the nurse licensing examination, the National Council of State Boards of Nursing (NCSBN) orchestrated a multi-year educational program to acquaint candidates with the CBT format and the testing model (CAT).

Implementation of a Centralized Examination Administration Call Center

Discussions with sponsors of credentialing programs delivering examinations in a CBT environment have highlighted the need to develop a centralized call center to respond to requests for information from state boards of accountancy as well as from candidates. The proposed centralized call center is designed to provide information about computer-based test administration and scoring to eligible candidates and board personnel. For example, in the case of candidates, center personnel might respond to questions about appointments for CBT administration, complaints about testing conditions, and requests for score transfers. In the case of boards of accountancy, center personnel might respond to questions about score interpretation, data transfers, reporting, and documentation of results.

The proposed centralized call center will reduce the volume and timing of calls received by state boards of accountancy concerning test administration issues, but will not impact the volume of calls concerning eligibility.

Development of a Comprehensive Database and Software for Boards of Accountancy and Other Key Stakeholders

Development of a single comprehensive database is necessary to ensure uniform data entry of candidate information, upload and download of candidate information, and the tracking of candidates. The database software should be provided to each state board of accountancy, the AICPA, NASBA, and any third party vendors involved in a subcontracting relationship with the
AICPA. Extensive quality control procedures are required to ensure the accurate and timely transfer of information among the AICPA, NASBA, the state boards of accountancy, and all third party vendors involved in a subcontracting relationship with the AICPA.

**Costs**

To develop cost estimates for the ideal examination, all expenses were classified into two major categories—development and implementation. Development costs include one-time costs associated with the activities necessary to build the program in order to implement the examination in March 2003. Moreover, development costs comprise costs which are specific to either Day 1 or Day 2 examination, as well as general development costs related to necessary system supports and infrastructures. Development costs are estimated conservatively at $9,720,000 (see Table 1 and Table 1 Supporting Schedule in Appendix 6, pages 68 – 69).

Implementation costs include all recurring costs required for the ongoing administration of the program during the first and subsequent years of testing. Implementation costs may vary by year, but are estimated on an annual basis in the form of per candidate costs for the Day 1 and Day 2 examinations. PES has based the per candidate cost on the assumption that each year 100,000 candidates will take a Day 1 and a Day 2 examination. The Day 1 per candidate cost is estimated to be $178; the Day 2 per candidate cost is estimated to be $198 (see Table 2 and Table 2 Supporting Schedules in Appendix 6, pages 70 – 72).

In compiling these costs, PES assumed that the AICPA would be responsible for all activities with the exception of explicitly subcontracted tasks. The labor and inflationary rates applied were agreed upon with AICPA. All salaries include an estimated base salary at January 1999 rates, plus 20% for benefits. A compounded 4% per annum inflation factor was applied to annual salaries starting in the year 2000.

**Development Costs**

General development costs. Three sets of general development costs were identified: for call center development, for database and software development for state boards of accountancy, and for educational materials development and duplication. General development costs are estimated at $1,410,000 plus overhead (see Table 1 and Table 1 Supporting Schedule in Appendix 6, pages 68 – 69).

The first general development project is the development of a centralized examination administration call center to support all required communications between the sponsors of the credentialing program, and the key stakeholders such as the candidates, the state boards of accountancy, and the practitioners. The estimated cost of this general development project, including equipment, hardware and software, system development, programming, and training of customer service representatives is $100,000 plus overhead. During implementation, it is anticipated that the center will be staffed by one full-time customer service manager and six full-time customer service clerks.
The second general development project is the development of a database and software for use by the AICPA, NASBA, all boards of accountancy, and any third party vendors involved in a subcontracting relationship with the AICPA. Development of a database is necessary to ensure uniform data entry of candidate information, uploading and downloading of candidate information, and the tracking of candidates. The database software will be provided to each state board of accountancy, the AICPA, NASBA, and any third party vendors involved in a subcontracting relationship with the AICPA. It is assumed that this task will be subcontracted; the estimated cost of this general development project is $700,000 plus overhead.

The third general development project is the development of educational materials necessary to acquaint the candidates with the new format and content of the ideal examination, including the use of a CBT delivery system and new item types. This project includes the development and production of two CD-ROM disks as well as printed educational materials. It is assumed that this task will be subcontracted; the estimated cost of this general development project is $610,000 plus overhead.

Development of Day 1 examination. Three sets of development costs were identified in connection with the development of the Day 1 examination: costs related to the implementation of new test specifications, item development, and form development. Total Day 1 development costs are estimated at $2,110,000 plus overhead (see Table 1 and Table 1 Supporting Schedule in Appendix 6, pages 68 – 69).

Day 1 development costs for the implementation of the new test specifications, including labor for three professional staff members, were estimated. Projected activities include classification of new items and reclassification of existing items. These development costs are estimated at $35,000 plus overhead.

In order to produce 24 sets of base tests and 24 sets of testlets with no more than 20% overlap from form to form, an item bank of no less than 4200 pretested calibrated items spanning all content areas must be available. PES understands that approximately 2000 items are available in the existing AICPA item bank. Therefore, for the development period, cost estimates were based on the development of 2200 additional objective items.

In order to project the overall costs for item development, PES used a per item development cost. As agreed upon with the AICPA, and on the basis of actual costs for the 1998 AICPA testing cycle, PES used a per item development cost of $849. This per item development cost included all labor and resources associated with item development. The item development costs are estimated at $2,021,000 plus overhead.

Day 1 development costs for examination form development, including labor, were estimated at $54,000 plus overhead.

Development of Day 2 examination. An overall cost was established in connection with all activities necessary to develop the Day 2 simulations. Total Day 2 development costs are
estimated at $5,500,000 plus overhead (see Table 1 and Table 1 Supporting Schedule in Appendix 6, pages 68 – 69).

The costs for the development of simulations for Day 2 testing were estimated on the basis of information obtained from at least four national credentialing organizations which have undertaken complex simulation development initiatives. These projects included extensive development phases, and are comparable in overall complexity to the projected parameters of the ideal examination. Additionally, the cost estimate for Day 2 is based on the assumption that the results of AICPA’s pilot project regarding the development and scoring of the simulations will provide support for the projected development and implementation schedule. Should additional cycles of examination development, review and revision be required, the cost will increase.

The types of development activities covered in connection with simulation development are prototype development, including design of simulation problems and scoring; acquisition of computational tools, including spreadsheet software and present value table; programming of prototype problems, including integration of computational tools and access to authoritative literature; expert reviews; prototype revision and debugging; expert re-reviews; pilot-testing; and beta-testing.

PES estimates that at least three complete revision cycles for the simulation problems will be required. Each cycle will include review and revision of the problem design and content. The scoring system will be evaluated and alternate scoring schemes will be developed for pilot testing. Standard setting will be conducted and the application of the standard will be evaluated. Psychometric and statistical analyses including analyses of problem functioning and reliability will be conducted. Empirical research will be conducted regarding the utility of automated scoring programs for the assessment of writing skills. Based on the results of these types of review, modifications will be implemented, and the modified problems will be pilot tested. When the problem design is finalized, “template” systems will be developed to facilitate future item development.

Implementation Costs

The total per candidate cost for Day 1 and Day 2 of the ideal examination is $376 (see Table 2 and Table 2 Supporting Schedules in Appendix 6, pages 70 – 72).

Per candidate cost for Day 1. The per candidate cost for Day 1 is estimated at $178 (see Table 2, Table 2 Supporting Schedules in Appendix 6). The estimate is based on testing 100,000 candidates annually. The examination will be developed by the AICPA and administered on a single computer network. CBT delivery costs have been estimated based on a mid-range of costs obtained from current CBT vendors using fixed-site systems. To estimate the CBT vendor administration cost, PES used the standards for secure delivery implemented by other high-stakes licensing examination programs administered on computer.
The components of the cost include examination development, CBT vendor administration for an 8-hour test day, scoring and all associated authentications and score transfer to boards, program management, duplication and development of revised educational materials, and travel and related expenses of committees and staff to attend meeting and related conferences.

**Per candidate cost for Day 2.** The per candidate cost for Day 1 is estimated at $198 (see Table 2 and Table 2 Supporting Schedules in Appendix 6). The estimate is based on testing 100,000 candidates annually. The examination will be developed by the AICPA and administered on a single computer network. CBT delivery costs have been estimated based on a mid-range of costs obtaining from current CBT vendors using fixed-site systems. To estimate the CBT vendor administration cost, PES used the standards for secure delivery implemented by other high-stakes licensing examination programs administered on computer.

The components of the cost include simulation development, template development, form development, CBT administration for a 6.5 hour test day, machine scoring of free-response tasks with subject-matter expert judgment of discrepant scores, customer services, program management, development of educational materials, and travel and related expenses of committees and staff to attend meetings and related conferences.

**Benchmarking Costs for the Development of a Linear Examination**

For purposes of comparing the cost of a CMT examination with the cost of a non-CMT examination, separate cost estimates were developed for a 1-day (i.e., 7 hour), multiple-choice, fixed-length linear examination, comparable in content coverage to Day 1 of the ideal examination. The per candidate cost included costs only for annual, ongoing item and form development, CBT administration, scoring, standard setting, program management, and travel. The per candidate cost for the linear examination is estimated at $179 (see Table 3 and Table 3 Supporting Schedule, pages 73 – 74). Analysis and comparison of these virtually identical cost estimates for Day 1 of the ideal examination and for the linear examination indicate that the major cost driver for Day 1 is the number of hours of testing time rather than the underlying testing model.
SUMMARY

Summaries of technical and stakeholder perspectives regarding the proposed ideal examination are provided, as well as an evaluation of the feasibility of developing and implementing the examination in terms of time and cost. PES has identified required activities to facilitate implementation, regardless of testing model. Finally, PES proposes a process for the systematic review of alternate testing models and decision-making regarding the testing model to be implemented in the computerized CPA examination.

Technical and Stakeholder Concerns

Day 1: Development and implementation of the CMT model. Although a CMT testing model can be implemented, the use of a variable-length testing model has created problems for other high-stakes credentialing programs. Stakeholders, including both candidates and boards, may not understand CMT's psychometric underpinnings, the fact that valid pass/fail decisions can be based on the administration of different numbers of items to different candidates, and may question the adequacy of the content coverage and the use of examination results for diagnostic feedback to these candidates. Because of the application of stopping rules, that is, rules that determine whether candidates have passed or failed the examination after the administration of only a base test and certain subsequent testlets, many candidates may fail a shortened test.

Day 2: Development and implementation of simulations. While simulations have excellent face validity because they closely mirror what practitioners do in their work, research is required to determine whether the scores they produce will be valid and reliable reflections of the knowledge and skills they are intended to measure.

PES identified several specific technical and psychometric concerns: content validity and the potential for inadequate coverage of the CSOs; unintentional effects that performance in one section may have on performance in another section of a simulation; the overall complexity of the simulations which require shifting among many computer screens and the use of on-line supplementary materials; the highly memorable content and associated security problems; the reliability of the simulations due to the limited number of scorable responses; feasibility of setting the passing standard and equating; and difficulty in developing adequate diagnostic feedback to failing candidates.

Implementation of the comprehensive 2-day examination. In the ideal examination, performances on Day 1 and Day 2 are equally weighted. While the Day 1 examination can produce valid and reliable scores meeting established psychometric criteria, the Day 2 examination, including only four 90-minute simulations, may not produce scores meeting defensible psychometric criteria in regard to reliability and validity. Further, the results of the proposed practice analysis and the associated CSOs must be carefully reviewed to evaluate the degree to which the contents of Day 1 and Day 2 complement each other and reflect the validated knowledge and skills base required for practice.
Market research is currently in progress regarding the implementation of the ideal examination. Preliminary feedback from key stakeholders indicates some initial misunderstanding regarding the description of the intended content and format of the ideal examination. Further research is required to evaluate key stakeholder support for the ideal examination.

Feasibility in Regard to Time Frames, Costs, and Implementation

The time frames for computerizing the ideal examination. Day 1 of the ideal examination can be implemented by 2003; however, the feasibility of implementing Day 2 by 2003 is somewhat problematic. The outcomes of the AICPA’s pilot project regarding the efficient development and scoring of valid and reliable simulations are required to properly evaluate this time frame.

The costs of developing and implementing the ideal examination. The ideal examination can be developed for $9.7 million, including the revision and implementation of the Content Specification Outlines (CSOs) subsequent to the conduct of the practice analysis; the development of educational materials and tutorials for candidates; the development of a single comprehensive database accessible to state boards and other key stakeholders; the development of items, forms, and simulations; computer programming for the development, administration, and scoring of the examination; the conduct of multiple cycles of reviewing and revising the examination; scoring; standard setting; equating; statistical and psychometric research and analysis; the selection of vendors; managing the program; and the implementation of quality control procedures throughout. The cost estimate for the ideal examination is based on the assumption that the results of AICPA’s pilot project regarding the development and scoring of the simulations provides support for the projected development and implementation schedule. Should additional cycles of examination development, review and revision be required, the cost will increase.

The per candidate cost for the ideal examination is estimated at $376, including $178 for Day 1 and $196 for Day 2. For purposes of comparison, costs were developed for a 1-day benchmark linear examination. The per candidate cost for a 7-hour linear examination is estimated at $179.

Required activities for implementation. PES has identified several significant events that must occur in order to computerize and implement the CPA examination, regardless of the testing model(s) selected. First, the AICPA must contractually obligate one or more CBT vendors to provide a sufficient number of testing sessions for the candidates. Second, the AICPA must identify security issues associated with the delivery of the examination, and contractually obligate its CBT vendors to minimize the likelihood of security problems occurring. Third, the current update practice analysis related to General Business Knowledge and the full-scope practice analysis must be carefully monitored, and the results incorporated into revised CSOs, so that future objective-item and simulation development efforts reflect contemporary practice. Fourth, as the absolute number of objectively scored items and simulations required for CBT is greater than the number on hand, the AICPA must continue to “ramp up” its test development.
efforts to produce and pilot-test the quantities of items and simulations required for ongoing implementation.

The AICPA must implement a research agenda to verify the facts and assumptions that PES made in regard to the ideal examination and/or any alternatives to that examination, because the accuracy of those facts and assumptions affected the conclusions reached regarding costs and timing. For example, the estimates PES made regarding the absolute number of candidates “in the pipeline” on an annual basis and the effects of using a compensatory scoring system for the Day 1 and Day 2 examinations and/or a non-compensatory system for combining the Day 1 and Day 2 examinations on the estimated pass/fail rates for either first-time or repeat candidates need to be verified through additional analyses. In addition, PES’s estimates of the frequency with which candidates might retake either day of the examination and the costs they would be willing to bear should be explored through further market research. Changes in the size of the candidate population will affect the numbers of items, testlets, forms, and simulations to be developed, pilot-tested, and published, in order to enhance examination security.

Recommendations

Establish financial parameters for initial examination development costs and ongoing operational costs. PES recommends that the Joint AICPA/NASBA CIC investigate the amount of money that various stakeholders are willing to pay in order to finance the initial examination development costs and the ongoing operational costs of the examination.

Continue public outreach. PES recommends that communication strategies continue to be enhanced to inform key stakeholders of the progress toward computerization and implementation of the ideal examination. Communication initiatives need to be developed and implemented to obtain stakeholder views and information as well as stakeholder concurrence.

Resolve potential problems. PES notes the following potential problems that need to be investigated and/or resolved prior to implementation: jurisdictions’ and candidates’ reactions to the price of the examination; uniform conditioning requirements across the jurisdictions; uniform transition policies across the jurisdictions; validating the form and content of the assessment of communications skills; and validating the utility of Day 2 for testing aspects of performance not addressed in an objective format.

Make policy decisions regarding the information flow to candidates and other stakeholders. PES recommends that the Joint AICPA/NASBA CIC make decisions regarding the

---

3 In a compensatory model, pass/fail decisions are a function of total test scores that are additive across all sections, content areas, and/or subtests within the examination. Low scores on one or more elements in the examination are compensated for by high scores on other elements. In a non-compensatory model, pass/fail decisions are a function of test scores on each section, content area, and/or subtest. Low scores on one element are not offset by high scores on a different element.
requirements for diagnostic feedback, the use of a pass/fail system versus scaled scores, and the use of composite scores versus subscores.

Establish a method for making a decision regarding the testing model to be implemented. Rather than endorsing the proposed CMT model, or suggesting an alternate model, PES believes that additional research must be conducted before a final determination is made regarding the testing model to be implemented. Accordingly, PES developed a method to facilitate the eventual selection of a testing model. The proposed method includes considering up to 24 features of testing models when comparing the models and making a final selection. Various psychometric-, stakeholder-, and cost-related drivers that interact with the features are identified and should be considered in evaluating the models.

Implementation of this method may lead to the elimination of several models from any further consideration. Then, using the comparison features, the strengths and limitations of the remaining models should be evaluated more closely. Further research may be required to make refined determinations regarding the interactions among the features and the remaining models as an aid to final decision-making.

Testing Models

Brief descriptions of six different testing models are provided. All of the models have both useful features and limitations, and may be used in a computerized environment.

**Linear.** This model consists of the computer administration of one or more fixed forms that are directly analogous to paper-and-pencil multiple-choice tests. Items may be administered in randomized order, but all candidates taking a form will see the same items. Forms may have overlapping items.

**Domain Sampling (DS).** This model is very similar to the linear model, except that the number of forms is much larger. The larger number of forms are made possible using automated test assembly procedures to construct multiple forms to the same content and statistical specifications. Forms are generally assembled in advance, rather than on the fly during the actual testing session, so that they may receive a traditional test form review by SMEs. Again, candidates taking a form will see the same items, but many fewer candidates will be expected to see each form. Forms may overlap, though the larger the item pool, the more limited the item overlap can be.

**Computer Adaptive Testing (CAT).** The test is adapted to each candidate. After an initial item is administered, candidates responding correctly to the first item will receive a harder item while those responding incorrectly will receive an easier one. Testing continues with each successive item chosen to collect the most information about the candidate’s ability. Administration may continue for a fixed number of items or may be ended once the estimate of the candidate’s ability is sufficiently precise to be confident about the score to be assigned.
Computer Mastery Testing (CMT). Short item sets (testlets) that are parallel in content and difficulty are randomly selected and successively administered until there is sufficient confidence about the status of a candidate as passing or failing according to a preset cutting point. In this model, the number of testlets administered is different for candidates of differing abilities, with the shortest test administered to those furthest from the cutting point. Scores for individuals are not provided by this testing method.

Computer Adaptive Mastery Testing (CAMT). In this model, items are administered as in CAT, but testing continues only until a decision can be made about the mastery status of the candidate, that is, passing or failing. The examinations typically have a minimum number of items that must be administered to all candidates, but otherwise are variable in length, with shorter examinations administered to candidates with abilities furthest from the preset cutting point. Scores for individuals are not provided by this testing method.

Computer Adaptive Sequential Testing (CAST). Short item sets (modules) are developed with parallel content, but different levels of difficulty. Testing is begun with a first-stage medium difficulty test and candidates are routed to an easier, more difficult, or (possibly) similar second-stage module depending on their performance on the first stage. Candidates would then be routed to a third-stage module in a similar fashion, depending on their performance on the second stage. The number of different stages and the number of different difficulty levels at a stage are part of the test design. Scores are provided.

Testing Model Features for Comparison

Twenty-four features associated with testing models are identified and organized within four topics—item and form development, administration, psychometrics/validity, and costs. These features can be used for highlighting the strengths and limitations of each testing model and for decision making regarding the selection of the final testing model.

1. Item and Form Development

Initial pool size. The item pool is the collection of items that are available for use in an examination. The items may be assembled into different forms, testlets, or modules, or may be made available in the computer for use in adaptive testing. In order to launch a computer-administered examination, a large number of items must be developed, often several times the number needed for a single linear form. The number of items required for the item pool will vary, depending on the testing model selected, the total needed for content coverage, the number of candidates, and the degree of concern for item security.

---

4 The following sections elaborate features previously identified by Parshall, Spray, Davey, & Kalohn, (1998).
In evaluating the interactions between the models and the feature, consider the following: Content specifications, statistical characteristics of the items in the bank, requirements for security, and candidate volume.

*Pool maintenance.* Maintenance refers to the need to replace items periodically once the program is underway, to keep content current and to prevent candidates from becoming familiar with item content. Programs that permit continuous testing will need more items than those that make use of administrative windows for testing. Testing models also differ in the degree to which large numbers of items need to be available at one time.

In evaluating the interactions between the models and the feature, consider the following: Rate of item exposure, periodic item recalibration, rate of change in field, and administration model.

*Pretest needed.* Pretesting items permits the detection of item flaws that are not caught in reviews and validates the correct answer with actual candidates. For all testing models, items will need to be pretested in order to develop test forms or to equate test forms. For some testing models, items will also need to be pretested in order to obtain data for the statistical analyses required for use of IRT, the statistical model that supports scoring examinations when candidates may have been given different items. This pretesting must be done prior to operational use of the items.

In evaluating the interactions between the models and the feature, consider the following: Program credibility, perception of fairness, and quality assurance.

*Pretest requirements.* Pretesting may be accomplished with any of the testing models by interspersing unscored pretest items with operational items. To obtain data for calibration, adequate candidate volume and sufficient numbers of previously used items are needed. Some methods may also require more items that provide information near the cutting score. Due to flawed item content or poor statistical properties, however, some items will be found to be inappropriate for use in future examinations. This probable loss must also be considered in planning for pretesting.

In evaluating the interactions between the models and the feature, consider the following: Item replacement rate needed to maintain pool, rate of pretesting loss, content specifications, statistical specifications, candidate volume, and administration model.

*Review of intact forms.* Reviews of the assembled form can reveal problems such as unfortunate combinations of items that may give away answers to one or more items or a lack of fit to the test specifications. Problems revealed by such reviews can then be corrected prior to administration. Such reviews can also be performed with individual testlets or modules prior to their publication for on-line administration. With CAT, however, items are selected by the computer at the time of testing, so that no prior review of the set of items to be taken by any one candidate is possible. On the other hand, automated test assembly procedures may be used to select items, while constraining the item selection to meet various complex criteria such as (a) achieving content balance, (b) keeping items together that must appear in sets, (c) matching
desired answer key distributions, (d) meeting equity concerns, and (e) avoiding overlapping items or items that should not appear together on a test form.

In evaluating the interactions between the models and the feature, consider the following: Test assembly procedures, perception of fairness, and quality assurance.

Diagnostic subscores. To form reliable subscores for diagnostic feedback for failing candidates, the number of items administered within each of the content areas of interest needs to be sufficient for this purpose. If subscores are desired, the testing design, including selection of the testing model, needs to take this desire into account. Variable-length tests may require additional constraints so as to ensure against too few items being administered for reliable diagnostic scores to be provided in all areas of the content specifications. With some models, testing may need to be continued beyond the point of making the pass/fail decision in order to perform effective diagnostic score reporting.

In evaluating the interactions between the models and the feature, consider the following: Content specifications; perceived fairness; and acceptability of program to candidates, boards, and practitioners.

Change of specifications. After a program is underway, the field being certified may change in such a way that modification of the specifications becomes necessary. With any testing model, changing specifications may require changing content codes that identify the items in the pool for manual or automated assembly or adaptive item selection. Changing specifications would also require changing the constraints for the automated assembly program or the item selection algorithms for adaptive testing. Changing the specifications may also require modification of forms, modules, or testlets already in use, which would then need to be republished. Extensive quality control may be required, regardless of the testing model.

In evaluating the interactions between the models and the feature, consider the following: Test validity, degree of change, databases, software, and quality assurance.

2. Administration

Overall testing time. Testing time must be reserved for the maximum time permitted for examinees even though many of them may leave early. Computer-based test delivery vendors may also require that adequate time be reserved for practice tutorials as well as the actual time spent testing. For testing sessions that go beyond half a day, the time provided to examinees for a meal may also become part of the time that must be scheduled. Hence, 6 hours of testing time may require at least 7 hours of seat time. Flexibility of scheduling testing time is increased if testing sessions are 4 hours or less.

In evaluating the interactions between the models and the feature, consider the following: Test delivery vendor; administration model; measurement precision; requirement for diagnostic subscores; and acceptability of program to candidates, boards, and practitioners.
**Fixed versus variable length.** In a fixed-length test, every examinee takes the same number of items, assuming that speed is not a factor. Linear, domain sampling, and CAST are fixed-length testing models. A variable-length test permits an examinee to stop testing once a score can be assigned or passing status determined. CAMT and CMT are variable-length testing models. CAT may be designed to have either fixed or variable length. Variable-length testing has been criticized in that failing candidates may feel that they have been “cut off” too quickly. In fact, testing may continue for the purpose of ensuring sufficient data for diagnostic reporting and adequate content coverage.

In evaluating the interactions between the models and the feature, consider the following: Administration model; perception of fairness; and acceptability of program to candidates, boards, and practitioners.

**Candidate item review.** When working through the examinations, candidates like to be able to return to previously seen items to review their responses. Linear and domain sampling forms permit candidates to review items at any time. With CMT testlets and CAST modules, items may be reviewed until candidates proceed to the next testlet/module. CAT and CAMT do not generally permit item review.

In evaluating the interactions between the models and the feature, consider the following: Perceptions of fairness, and acceptability of program to candidates, boards, and practitioners.

**Item exposure.** When a candidate views a test item, it is exposed. For security purposes, increasing the number of times that items are exposed can become a serious security issue, since greater exposure increases the probability that the item will become compromised. Although security may be better with computer-based testing than with paper-and-pencil testing, instances have occurred in the past to make the possibility of item compromise a concern for programs. In general, increasing the number of items available for use will reduce exposure, although item exposure will also vary according to the testing model selected and the safeguards against exposure that have been adopted. With certain models, conditional item exposure rates (conditioned on ability level) need to be monitored very carefully.

In evaluating the interactions between the models and the feature, consider the following: Pool size, pool maintenance, administration model, test score validity, security, and candidate volume.

**Item challenges.** Items may be challenged by examinees who believe that an item is flawed or who want their scores confirmed. (The fact that items may be challenged supports the general policy of not providing feedback to candidates at the time of the examination.) Legal challenges to test items occur most often from failing candidates. With variable-length tests, these candidates may feel that they have not had a fair chance to demonstrate their abilities, particularly when testing has ended after relatively few items. Their feeling may be that they could have passed if they had had an opportunity to respond to more items.

In evaluating the interactions between the models and the feature, consider the following: Perceptions of fairness and quality assurance.
3. Psychometrics/Validity

Measurement precision. Measurement precision refers to the accuracy with which a test score approximates the "true" underlying ability or competence in the area being measured. Precision is generally expressed as the reliability or standard error of measurement of the scores. The theoretical advantage of CAT or CAST testing is that those items or modules are administered that improve measurement precision for the individual candidates across the ability range. The computer mastery models (CMT and CAMT) are designed to improve measurement precision specifically at the cut point.

In evaluating the interactions between the models and the feature, consider the following: Testing purpose, test length, and, statistical specifications.

Item selection algorithm. Item selection algorithms are the rules by which items are selected in a CAT administration. These algorithms must be in place in order to prevent over- and under-exposure of items. Research is required to identify the ideal starting point in an examination and the change in difficulty between items.

In evaluating the interactions between the models and the feature, consider the following: Item exposure and pool size.

Decision accuracy. The accuracy with which a test classifies candidates into passing and failing categories is referred to as decision accuracy. Decision accuracy is important in every testing model; it is dealt with explicitly in CAMT and CMT. In general, decision accuracy will be enhanced if the test is designed to yield more information in the ability region surrounding the cut score.

In evaluating the interactions between the models and the feature, consider the following: Testing purpose, statistical specifications, and test length.

Content coverage. Part of the validity of a credentialing examination lies in how well it covers the knowledge, skills, and abilities required for practice. Regardless of the test model, test length must be sufficient to make the case for content validity. In general, longer examinations may be perceived as doing a better job of covering content than shorter examinations.

In evaluating the interactions between the models and the feature, consider the following: Test validity; practice analysis; perceived fairness; and acceptability of program to candidate, boards, and practitioners.

Scale score continuity. Equating is required to ensure comparability of scores across forms. Candidate performance on the examination is typically transformed to a scale score to permit the reporting of comparable scores for candidates who have taken different forms of a test or who have taken different sets of test items. Continuity of the scale scores over time is important in order to maintain the integrity of the program. In the case of linear and DS models, raw score
reporting may be used; IRT-based score reporting may be required with CAT, CMT, CAST, and CAMT.

In evaluating the interactions between the models and the feature, consider the following: Test validity, perceived fairness, and quality assurance.

**Model robustness.** Robustness refers to the extent to which the assumptions of a statistical model can be violated and still produce adequate results. Although IRT provides the mathematical model underlying the different methods, the testing models differ in how accurate the model estimates need to be to provide adequately comparable scores. The linear and domain sampling models rely on IRT primarily to place the scores on the scale, while CAST and CMT depend on IRT for comparing and combining results from different testlets/modules. Since these applications primarily use summary information, poorly estimated statistics for a small number of items should not unduly affect the final outcome. CAT and CAMT, however, require that the IRT model fit each item and be calibrated accurately, since these statistics for individual items provide the underlying mechanism for item selection and generation of scores. In these models, poorly estimated items may affect outcomes. Moreover, violations of the single underlying trait in the item bank could be highly problematic for several of the test models.

In evaluating the interactions between the models and the feature, consider the following: Candidate volume and test length.

**Standard setting.** Standard setting is the procedure used to assign a cutting point on the score scale. For credentialing examinations, standards are most often set using a method by which judges evaluate the individual items on the examination and the probable performance by examinees whose abilities are just adequate to be considered competent. Although these judgmental procedures can be used with any of the models, the standard will be most directly translated from the study results to the ability scale when intact forms, testlets, or modules can be evaluated. With CAST, the items in one or more complete pathways can also be evaluated for standard setting. For adaptive testing, it is less clear which items should be evaluated for judgmental methods to be used.

In evaluating the interactions between the models and the feature, consider the following: Test validity and perception of fairness.

4. Costs

**Item development.** The development of the item pool for computerized testing and the ongoing development of new items to keep the item pool refreshed are likely to be the largest expenses of the program, regardless of model. However, some models require larger item pools initially and, depending on item exposure, will require more items to keep them functioning as desired. The importance of item exposure to maintenance of security is key to determining costs for upkeep of a program. If item exposure is a serious concern, more items will need to be developed. If it is of lesser concern, fewer items will be needed to maintain the program. The choice of the test model will also impact the statistical nature of the items to be included in the
pool. For example, some models may require more items near the cut point, but such items may be difficult to produce, thus requiring more initial pretesting of items.

In evaluating the interactions between the models and the feature, consider the following: Statistical specifications, content specifications, security, item development procedures, the use of new item formats, and candidate volume.

**Pretesting.** In all of the models, unscored pretest items can be administered interspersed among the operational items. Data from pretesting can be used to verify the appropriateness of items, to obtain statistics to make pretest items available for operational use, and to perform differential item functioning analyses.

In evaluating the interactions between the models and the feature, consider the following: Item replacement rate needed to maintain pool, rate of pretesting loss, content specifications, statistical specifications, candidate volume, and administration model.

**Form development.** Once the program is underway, new forms/testlets/modules can be created from previously used pool items and pretested items. New forms are not needed with adaptive testing; rather, pretested items are entered into the active item pool as they become available and high exposure items are retired.

In evaluating the interactions between the models and the feature, consider the following: Security, administrative model, candidate volume, and software requirements.

**System development.** Linear and CAT testing have commercially available drivers that can be used by clients of different computer-based test delivery providers. CAMT and CMT are available through the Educational Testing Service, though other providers would probably work with AICPA to develop these capabilities, because they are not very different from current capabilities. Domain sampling would require no special driver, but providers may charge extra fees to keep many full-length examinations available. CAST does not yet have a fully operational system for computer administration.

In evaluating the interactions between the models and the feature, consider the following: Test delivery vendor and software requirements.

**Candidate education.** Candidate education and other outreach activities are likely to become higher-cost items the more the model deviates from the paper-and-pencil mode with which candidates are familiar. Hence, candidate educational materials for linear and domain sampling will typically be relatively low cost because these models depart little from conventional testing. Candidates will require more education about the other testing models.

In evaluating the interactions between the models and the feature, consider the following: Perception of fairness and acceptability of program by candidates, boards, and practitioners.
## Appendix 1
### Psychometricians and Testing Construction Experts Interviewed and Discussion Topics

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Reason For Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>James H. Adair</td>
<td>Lotus Development Corporation</td>
<td>Use of performance assessments and complex item types in certification setting. Availability of software.</td>
</tr>
<tr>
<td>Randy E. Bennett</td>
<td>Educational Testing Service</td>
<td>Research with computerized scoring of complex performance assessments and constructed response items.</td>
</tr>
<tr>
<td>Betty A. Bergstrom</td>
<td>Computer Adaptive Technologies, Inc</td>
<td>Experience with computer-based testing. Representative of vendor for computer delivery of tests.</td>
</tr>
<tr>
<td>Anna Bersky</td>
<td>National Council State Boards of Nursing</td>
<td>Worked on development of computer-based simulations for nursing.</td>
</tr>
<tr>
<td>John Boyce</td>
<td>National Board Examination Committee for Veterinary Medicine</td>
<td>Planning implementation of computer-based examination.</td>
</tr>
<tr>
<td>Jill Burstein</td>
<td>Educational Testing Service</td>
<td>Member of team developing E-Rater, computerized essay scoring system.</td>
</tr>
<tr>
<td>Brian Clauser</td>
<td>National Board of Medical Examiners</td>
<td>Research on automated scoring and standard setting for computer-based case simulations and other performance assessments.</td>
</tr>
<tr>
<td>Steve G. Clyman</td>
<td>National Board of Medical Examiners</td>
<td>Led development of computer-based case simulations in medical area. Worked with nurses in development of their simulations.</td>
</tr>
<tr>
<td>Fritz Drasgow</td>
<td>University of Illinois</td>
<td>Developed computer-based simulations for use in industry.</td>
</tr>
<tr>
<td>Robin Durso</td>
<td>Educational Testing Service</td>
<td>Worked on development of computer-adaptive Graduate Record Examination, an early implementation of computer-based testing.</td>
</tr>
<tr>
<td>Name</td>
<td>Organization</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Scott Elliott</td>
<td>Vantage Technologies</td>
<td>Developed Intellimetric product for scoring essays by computer.</td>
</tr>
<tr>
<td>Michael Englander</td>
<td>Uniscore</td>
<td>Provides handscoring services for computer-administered essays.</td>
</tr>
<tr>
<td>David F. Foster</td>
<td>Galton Technologies</td>
<td>Experienced with computer-based testing, primary in IT industry.</td>
</tr>
<tr>
<td>Ronald K. Hambleton</td>
<td>University of Massachusetts</td>
<td>AICPA External Consultant. Research on computer-based testing and standard setting with performance assessments.</td>
</tr>
<tr>
<td>Ron Hanscher</td>
<td>National Associations of Boards of Pharmacy</td>
<td>Implemented computer-adaptive test for pharmacy exams.</td>
</tr>
<tr>
<td>Ellen R. Julian</td>
<td>American Association of Medical Colleges</td>
<td>Formerly at the National Council State Boards of Nursing, where she was in charge of development of computer-administered licensing examinations using computer adaptive mastery testing.</td>
</tr>
<tr>
<td>Michael T. Kane</td>
<td>University of Wisconsin</td>
<td>NASBA External Consultant. Research on standard setting.</td>
</tr>
<tr>
<td>Jeffrey Kenney</td>
<td>National Council of Architectural Boards</td>
<td>Led development effort for Architects examinations including examinations using both computer mastery testing (testlets) and simulations. Presently an independent consultant, Professional Development Partners, Inc.</td>
</tr>
<tr>
<td>Charles Lewis</td>
<td>Educational Testing Service</td>
<td>Research for psychometric underpinnings of computer mastery testing (testlets).</td>
</tr>
<tr>
<td>Richard M. Luecht</td>
<td>National Board of Medical Examiners</td>
<td>Research and development for Computer adaptive sequential testing (CAST), using testlets, and automated test assembly. Presently at the University of North Carolina—Greensboro.</td>
</tr>
<tr>
<td>Robert Mislevy</td>
<td>Educational Testing Service</td>
<td>Development of simulations in dental areas.</td>
</tr>
<tr>
<td>Larry Newman</td>
<td>Assessment Systems Inc.</td>
<td>Experience with computer-based testing. Representative of vendor for computer delivery of tests.</td>
</tr>
<tr>
<td>Barbara S. Plake</td>
<td>University of Nebraska—Lincoln</td>
<td>Research in computer-based testing and in standard setting for performance assessments.</td>
</tr>
<tr>
<td>Name</td>
<td>Organization</td>
<td>Experience/Role</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Janice D. Scheuneman</td>
<td>Quality Assessment Services</td>
<td>Research with differential item functioning, and computerization of paper and pencil tests.</td>
</tr>
<tr>
<td>Judith Spray</td>
<td>ACT</td>
<td>Research and experience with computer-based testing. Development work on automated test assembly (domain sampling).</td>
</tr>
<tr>
<td>C. David Vale</td>
<td>The Chauncey Group International</td>
<td>Experience with computer-based testing.</td>
</tr>
<tr>
<td>Anthony Zara</td>
<td>National Council State Boards of Nursing</td>
<td>Currently leads psychometric work for nursing examination using computer adaptive mastery testing. Research on computer-based testing.</td>
</tr>
</tbody>
</table>
Appendix 2
Exemplars of Related Research Literature


## Appendix 3
### CBT Vendors

<table>
<thead>
<tr>
<th>CBT Vendor</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment Systems, Inc.</td>
<td>Bala Cynwyd, Pennsylvania</td>
</tr>
<tr>
<td>Sylvan Prometric</td>
<td>Baltimore, Maryland</td>
</tr>
<tr>
<td>Cogent Testing Network (Experior)</td>
<td>St. Paul, Minnesota</td>
</tr>
<tr>
<td>National Computer Systems, Inc./ Virtual University Enterprises</td>
<td>Eden Prairie, Minnesota</td>
</tr>
</tbody>
</table>
Appendix 4
Credentialing Agencies and Information Technology Companies
Implementing CBT Programs

American Association of Critical Care Nurses (certification of critical care nurses)
APICS (formally, The Educational Society for Resource Management) (certification of inventory control professionals)
Board for Certification of Emergency Nurses (certification of emergency nurses)
Certification Board for Perioperative Nursing (certification of perioperative nurses & RN first assistants)
Dental Interactive Simulation Corporation (proposed CBT assessment of dental technologists)
Federation of State Boards of Physical Therapy (licensing of physical therapists and physical therapy assistants)
Hewlett-Packard Company (certification of software professionals)
Lotus Development Corporation (certification of software professionals)
Microsoft Corporation (certification of software professionals)
National Association of Boards of Pharmacy (licensing of pharmacists)
National Board of Medical Examiners (licensing of physicians)
National Board of Podiatric Medical Examiners (licensing of podiatrists)
National Council of Architectural Registration Boards (licensing of architects)
National Council of State Boards of Nursing (licensing of nurses)
Appendix 5
Details Regarding the Facts and Assumptions

• General assumptions regarding the administration of Day 1 and Day 2 of the examination

1. Day 1 of testing will include no more than 8 hours of testing time. This includes the administration of a tutorial of up to 15 minutes in length and the administration of a set of pilot-test items and live items.

2. Day 2 of testing will include no more than 6.5 hours of testing time. This includes the administration of a tutorial of up to 30 minutes in length and the administration of 4 simulations.

3. First-time candidates (and repeat candidates not previously passing any day of the examination) will be required to register for both Day 1 and Day 2 of the examination.

Comment: PES understands that all but five jurisdictions require that candidates register for and take all sections of the examination not previously passed. Potentially, allowing candidates to register for a reduced number of sections permitted them to focus their study. Since both Day 1 and Day 2 of the ideal examination include comprehensive coverage of content, there is no benefit to registering for less than the full examination. Accordingly, to enhance the implementation of the computerized testing program, PES recommends that candidates “begin” the process in a uniform manner.

4. Candidates will be permitted to schedule the administration of Day 1 and Day 2 in any order.

Comment: Day 1 and Day 2 are both required. Permitting candidates to schedule Day 1 and Day 2 in the sequence they prefer may be seen by the candidates as a “positive.”

5. Conditioning requirements will be uniform across the jurisdictions.

Comment: PES understands that this is a policy issue not yet settled; however, uniform requirements would enhance the implementation of the computerized testing program. Moreover, the adoption of uniform requirements is consistent with the general trend in licensed professions to facilitate interstate mobility. Finally, given that the ideal examination encompasses the equivalent of only two sections, the variations on conditioning are more limited than at present, e.g., the length of time “pass” decisions on Day 1 or Day 2 may be banked; the number of retries a candidate will be permitted.

6. Conditioning requirements will be developed such that candidates passing Day 1 or Day 2 will be permitted up to 5 retries on the examination not previously passed within a 3-year period.
Comment: The only relevant aspects of current conditioning requirements which might be applied to the ideal examination relate to the length of time scores can be “banked” and the number of retries a candidate is permitted. Credentialing programs may decide to control the number of retries a candidate is permitted in order to reduce the chances of false positive decisions being made with regard to borderline candidates. Similarly, credentialing programs may require failing candidates to re-initiate their period of eligibility pending additional education or preparation for the examination.

7. Candidates will be permitted to register and test for both Day 1 and/or Day 2 of the examination up to four times within a 12-month period, but no more frequently than once in each 3-month period.

Comment: One major advantage of CBT is the increased frequency with which candidates may sit for an examination. Currently, candidates may sit for the examination twice a year; accordingly, four administrations per year would represent a significant increase in the frequency at which the examination is offered.

8. Four testing windows will be established annually. Each window will be approximately 60 consecutive days, including 48 testing days (6 days per week, Monday to Saturday, inclusive), followed by approximately 30 consecutive non-testing days.

Comment: The demand for testing may not be consistent throughout the year; however, it is not possible to predict what the periods of peak demand will be. The experiences of other credentialing programs suggest that candidates may delay the initial scheduling of their examinations rather than use the opportunity to schedule their first examination immediately upon becoming eligible to test.

In general, the use of testing windows permits jurisdictions to schedule and/or control work flow. Typically, boards establish specific time frames wherein: (a) applications are processed, (b) candidate eligibility is determined, and (c) score reports are processed and candidates are notified of their results. Separately, the 30 non-testing days in each window are used for the conduct of CBT-vendor administrative functions such as: (a) publication of new test forms, and (b) beta-testing of new test forms.

9. All candidates registering for both Day 1 and Day 2 shall be scheduled to test within the same testing window. The CBT vendor shall schedule all candidates for the examination within 30 days of their desired test dates. In order to guarantee seat time for candidates, the AICPA must contractually obligate the CBT vendor to provide adequate access to candidates.

10. Retake candidates who have passed Day 1 or Day 2 will be permitted to register for the one day not previously passed.

11. Between 80,000 and 100,000 different candidates are “in the system” during each 12-month period. PES believes that fewer than 120,000 different candidates test annually.
Approximately 40,000 *first-time candidates* register for the examination each year, and fewer than 80,000 *different repeat candidates* take the examination each year.

Comment: This assumption is based on information contained in the document *Candidate Performance on the Uniform CPA Examination, 1998 edition* (NASBA), indicating that (a) the same repeat candidates may sit for both the May and November examinations, and (b) five jurisdictions permit candidates to take fewer than all of the subjects for which they are eligible and receive grades only in those subjects completed, and subsequently identify those candidates as repeat candidates.

- **General assumptions regarding the scoring of the Day 1 and Day 2 examinations**

1. Candidates will receive separate pass/fail scores for Day 1 and for Day 2.

2. Failing candidates will receive separate diagnostic information for Day 1 and for Day 2.

   Comment: *The Standards for Educational and Psychological Testing* (AREA, NCME, APA, 1985) indicate that candidates should be provided diagnostic information regarding their performance on credentialing examinations. Although the Standards are currently undergoing revision, publicly disseminated drafts continue to include the requirement for diagnostic information.

3. Candidates will not receive scores on-site.

   Comment: High-stakes credentialing programs have generally not supported the on-site release of scores. Following each CBT administration, separate quality assurance procedures may be implemented by the CBT vendor and by the scoring vendor. Additionally, Day 2 of the ideal examination may require external scoring by experts in regard to the essays. Finally, some credentialing agencies have expressed concern over the possible reactions of candidates when confronted with “failing” scores.

4. Candidates will be permitted to “carry over” examination scores for no more than 3 years or 5 tries, whichever limit is reached first.

   Comment: Credentialing programs may decide to control the number of retries a candidate is permitted in order to reduce the chances of false positive decisions being made with regard to borderline candidates. Similarly, credentialing programs may require failing candidates to re-initiate their period of eligibility pending additional education or preparation for the examination.

5. The passing rates for Day 1 and Day 2 will be equal for first-time candidates. On the basis of a review of current candidate performance, more than 25% of first-time candidates should pass Day 1 and more than 25% of first-time candidates should pass Day 2. A subset of these first-time candidates should pass both Day 1 and Day 2, and
fewer than 60% of first-time candidates will fail both Day 1 and Day 2. Moreover, the percentage of first-time passing candidates may increase as a function of (a) the compensatory scoring model to be implemented, and (b) implementation of the 5-year degree requirement.

PES understands that passing rates are generally lower for repeat candidates than they are for first-time candidates. The current assumptions do not take into account any differences between the groups of first-time candidates and repeat candidates.

Comment: PES is aware that the AICPA is currently undertaking research to evaluate these assumptions and refine the pass/fail estimates.

- Specific facts and assumptions regarding Day 1 of the ideal examination

1. All items are objective; item types will include graphically enhanced, multiple-choice questions (MCQs) and a variety of drop-and-drag questions, matching questions, and other selected-response questions. Some items may be interactive.

2. Testing time is estimated at 108 seconds per objective item, including MCQs and other objective formats (OOFs).

Verification of the time required to complete objective items in a CBT environment is required to refine the cost estimates for Day 1.

3. The testing model for Day 1 is CMT including the administration of a base test, and between 1 and 6 parallel testlets per candidate, if necessary.

4. All candidates will take a morning testing session of not more than 3.3 hours, plus a 15 minute tutorial period. Candidates for whom a pass/fail decision has not been reached will take an afternoon session of not more than 4.5 hours.

5. Each candidate is to take a morning session including the base form consisting of 100 live items and 10 pilot-test items.

Comment: Whereas base forms may be as short as 60 items, PES endorses the use of a base form including 100 items, both to ensure content coverage and to be responsive to stakeholder concerns that the examination appear to be rigorous.

6. Candidates are not given a formal rest break during the morning session, but are permitted to take rest breaks as necessary.

7. Each candidate requiring an afternoon testing session will take 1, 2, 3, 4, 5, or 6 parallel testlets, consisting of 25 items each (2 testlets = 1.5 hours; 4 testlets = 3 hours; 6 testlets = 4.5 hours) for a maximum afternoon testing session of 4.5 hours.
8. Performance estimates based on the morning base test will produce a pass/fail decision for as many as 50% of the candidates.

9. Performance estimates based on the afternoon testlets will produce a pass/fail decision for as many as an additional 10% of the candidates after 2 testlets, an additional 15% of the candidates after 4 testlets, and the remaining 25% of the candidates after 6 testlets.

• Specific assumptions regarding Day 2 of the ideal examination

1. During Day 2, candidates will be administered 4 simulations in two 3-hour testing sessions.

2. Each simulation is a separate event lasting no more than 90 minutes; candidates may not return to a simulation once they have completed the simulation.

3. The morning session is no more than 3.5 hours, including a tutorial requiring no more than 30 minutes, and a 3-hour testing period. The afternoon session is no more than 3 hours.

Comment: PES has recommended that practice materials be developed, including a CD-ROM. The practice materials will be designed to familiarize candidates with the form and content of the objectively scored items and of simulations, including the use of research databases, spreadsheets, and CAT software, and the expected narrative and analytic responses.

4. Pilot-testing of simulations will not occur as part of operational testing.

Comment: PES understands that during the operational phase of the ideal examination, pilot-testing may be conducted with students, volunteers, and recently certified CPAs.

5. Each simulation will be constructed to assess knowledge and skill related to more than one of the four sections of the current examination. The work of the Content Oversight Task Force (COTF) in regard to the development of test specifications for the simulations, as well as the results of all recently completed practice analysis studies, will provide guidance regarding the development of the test specifications for the simulations.

6. Each simulation will include a text- and graphics-based scenario. Each simulation will require the use of custom-developed on-line software such as spreadsheets, research databases, CAT software to audit in an information technology environment, authoritative literature such as generally accepted accounting principles (GAAP) and generally accepted auditing standards (GAAS), and/or IRS laws and regulations.
7. Candidates will be required to demonstrate research and/or problem solving skills as well as communications skills in each simulation.

8. Communications skills tested within each simulation will include writing skills such as bullet point responses, short essays including 3 to 5 concepts, and long essays including 6 to 10 concepts. Communications skills may also include the production of graphic and tabular presentations.

9. Performance on simulations may be scored for process and content.

10. In spite of attempts to reduce interdependence, there may be interdependence among the sections within each simulation.

11. Two approaches will be used to score written responses, live readers and automated scoring. A combination of approaches may be used, e.g., an initial screening via automated scoring, and the use of readers for borderline responses.

• Specific assumptions regarding the development and administration of a linear examination

1. The linear examination consists of objectively scored items, including graphically enhanced MCQs and other selected-response questions.

2. The non-MCQs are estimated to comprise 30% of the total number of objectively scored items.

3. Costs for development and administration will be presented on the basis of the following scenario: 7 hours of testing including a 15 minute tutorial, and a testing period including live items and pilot-test items.

4. Each candidate’s test will be completed in one day.

5. Item development costs will be based on the current AICPA costs for objectively scored questions.

6. CBT costs will be based on an average of the current vendor charges for hourly seat time for programs of comparable candidate volume.

(All other relevant assumptions regarding Day 1 remain as described previously in connection with Day 1 of the ideal examination.)
**Assumptions regarding costing**

1. Research and development and programming costs necessary to develop simulations are presented separately from ongoing implementation costs. These costs are presented as ranges based on data gathered from other groups who have undertaken similar projects.

2. Costs for ongoing administration include costs for the CBT vendor (existing or newly developed network), test development, scoring and reporting, and overall management functions.

3. Software costs for a board and for the AICPA are included in cost estimates.

4. Capital expenses for establishing a testing center are not included.

5. Per item development costs for the objectively scored items are based on the current costs for these activities supplied by the AICPA.

6. Grading costs for expert-graded essays are based on the current costs for these activities supplied by the AICPA.

7. Computer-based administration costs are based on an average cost per hour for the current commercial vendors supplying CBT services for programs of comparable volume. In cases where costs vary by length of test, the variations in costs are outlined. Day 1 is costed on the basis of an 8-hour administration; Day 2 is costed on the basis of a 6.5-hour administration.

8. Current board of accountancy costs can be compared to board costs for examination administration in a CBT environment.

**Overall assumptions concerning the roles and responsibilities of the state boards of accountancy**

1. Applicants for licensure submit their initial application to state boards which will continue to determine candidate eligibility.

2. All state boards of accountancy will use the same software to enter, transmit, and receive data to be maintained in a centralized candidate database. Data on eligible candidates will be entered by each board and be transmitted electronically to the entity responsible for scheduling initial and subsequent testing sessions.

3. All state boards of accountancy will use specially designed software developed to support the centralized candidate database.
4. Each candidate will be assigned a unique identifying number which will remain constant. If the social security number cannot be used, another unique number will be assigned. All state boards will conform to this requirement.

5. Until eligibility is determined, candidates will communicate directly with the state board of accountancy. Following the granting of eligibility, the state board will send information to the candidates, regarding registration procedures. Candidates will then schedule their own examination testing sessions directly with the CBT vendor.

6. The AICPA (or other designated agency) will receive score data from the CBT vendor and may receive data from a second source in regard to the scoring of the essays. These data will be verified for accuracy of scores. Verified scores will be transmitted to state boards of accountancy.

7. Score data for individual candidates will be transmitted electronically by the AICPA (or other designated agency) to the state boards of accountancy.

8. All final licensure decisions will be made by the state boards of accountancy.

9. Conditioning requirements will not vary by jurisdiction.

10. Licensure certificates will be prepared and mailed to individual candidates by the licensure board.
Index

AICPA/NASBA Computer-Based Testing Feasibility Study
Cost Estimates

<table>
<thead>
<tr>
<th>Table</th>
<th>Page(s)</th>
<th>Page nos. in text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1 - Summary of development cost for ideal exam</td>
<td>68</td>
<td>30-32</td>
</tr>
<tr>
<td>Supporting schedule of costs for table 1</td>
<td>69</td>
<td>30-32</td>
</tr>
<tr>
<td>Table 2 - Summary of implementation costs for ideal exam</td>
<td>70</td>
<td>32-33</td>
</tr>
<tr>
<td>Supporting schedules of costs for table 2</td>
<td>71-72</td>
<td>32-33</td>
</tr>
<tr>
<td>Table 3 - Summary of implementation costs for linear exam</td>
<td>73</td>
<td>33-34</td>
</tr>
<tr>
<td>Supporting schedule of costs for table 3</td>
<td>74</td>
<td>33-34</td>
</tr>
</tbody>
</table>
## Table 1- Development of Ideal Exam

### General Development (1)

<table>
<thead>
<tr>
<th>Description</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development of call center</td>
<td>$100,000</td>
</tr>
<tr>
<td>Database development / Software for state boards</td>
<td>$700,000</td>
</tr>
<tr>
<td>Educational materials development</td>
<td>$250,000</td>
</tr>
<tr>
<td>Educational materials duplication</td>
<td>$360,000</td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td><strong>$1,410,000</strong></td>
</tr>
</tbody>
</table>

### Development of Day 1 exam

<table>
<thead>
<tr>
<th>Description</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implementation of new test specifications</td>
<td>$35,000</td>
</tr>
<tr>
<td>Item development</td>
<td>$2,021,000</td>
</tr>
<tr>
<td>Examination development</td>
<td>$54,000</td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td><strong>$2,110,000</strong></td>
</tr>
</tbody>
</table>

### Development of Day 2 exam

<table>
<thead>
<tr>
<th>Description</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation development</td>
<td>$5,500,000</td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td><strong>$5,500,000</strong></td>
</tr>
</tbody>
</table>

### Overhead

<table>
<thead>
<tr>
<th>Description</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overhead</td>
<td>$700,000</td>
</tr>
<tr>
<td><strong>Grand Total - Ideal Exam</strong></td>
<td><strong>$9,720,000</strong></td>
</tr>
</tbody>
</table>

(1) See supporting schedule and footnotes on Page 69
### Table 1: Supporting Schedule

#### Ideal Examination - Development Costs

<table>
<thead>
<tr>
<th>Task</th>
<th>Inflation Factor (1)</th>
<th>Personnel</th>
<th>% of Time</th>
<th>Annual Salary (2)</th>
<th>Total</th>
<th>Direct Cost Per Unit</th>
<th>Estimated Unit Costs Per Unit</th>
<th>(Rounded) Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development of call center</td>
<td>4%</td>
<td>100,000</td>
<td>100,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop database &amp; software for state boards (3)</td>
<td>4%</td>
<td>700,000</td>
<td>700,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development of candidate educational materials (3)</td>
<td>4%</td>
<td>250,000</td>
<td>250,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Development and programming of two CD ROM's</td>
<td>4%</td>
<td>200,000</td>
<td>200,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Materials Production</strong></td>
<td></td>
<td>200,000</td>
<td>1.80</td>
<td>360,000</td>
<td></td>
<td>360,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-Total - General Costs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1,410,000</td>
<td>1,410,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Inflation Factor (1)</th>
<th>Personnel</th>
<th>% of Time</th>
<th>Annual Salary (2)</th>
<th>Total</th>
<th>Direct Cost Per Unit</th>
<th>Estimated Unit Costs Per Unit</th>
<th>(Rounded) Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Day 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation of new test specifications [2000]</td>
<td>4.0%</td>
<td>20%</td>
<td>96,000</td>
<td>19,968</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical Manager</td>
<td>4.0%</td>
<td>1</td>
<td>20%</td>
<td>96,000</td>
<td>19,968</td>
<td>20,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical programmer</td>
<td>4.0%</td>
<td>1</td>
<td>10%</td>
<td>90,000</td>
<td>9,000</td>
<td>9,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Editorial</td>
<td>4.0%</td>
<td>1</td>
<td>10%</td>
<td>60,000</td>
<td>6,000</td>
<td>6,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Development</td>
<td>8.2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form Development: [24 AM sets and 24 PM sets]</td>
<td>12.4%</td>
<td>50%</td>
<td>96,000</td>
<td>53,952</td>
<td>54,000</td>
<td>54,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program Director [2002]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-Total day 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>89,520</td>
<td>2,020,960 2,110,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Inflation Factor (1)</th>
<th>Personnel</th>
<th>% of Time</th>
<th>Annual Salary (2)</th>
<th>Total</th>
<th>Direct Cost Per Unit</th>
<th>Estimated Unit Costs Per Unit</th>
<th>(Rounded) Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Day 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation Development (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-Total day 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5,500,000</td>
<td>5,500,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Inflation Factor (1)</th>
<th>Personnel</th>
<th>% of Time</th>
<th>Annual Salary (2)</th>
<th>Total</th>
<th>Direct Cost Per Unit</th>
<th>Estimated Unit Costs Per Unit</th>
<th>(Rounded) Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overhead</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total general, day 1, and day 2 costs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9,720,000</td>
<td>9,720,000</td>
</tr>
</tbody>
</table>

---

(1) Inflation assumption: 4% per annum compounded starting in the year 2000. Inflation factors are applied to annual salaries and item development cost. Example: Salaries are estimated at January 1999 rates. 4% inflation is added for 2000, 8.2% for 2001, 12.4% for 2002, and 17% for 2003.

(2) Salaries include estimated base salary plus 20% for benefits.

(3) Estimated cost: to be sub-contracted by the AICPA.

(4) Based on an analysis of development costs obtained from four other national credentialing agencies which have undertaken simulation development similar to the projected parameters of day 2 testing.
### Table 2 - Implementation of Ideal Exam

<table>
<thead>
<tr>
<th>Ideal Exam</th>
<th>Day 1 (1)</th>
<th>Day 2 (2)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination development</td>
<td>$24.00</td>
<td>$25.00</td>
<td>$49.00</td>
</tr>
<tr>
<td>CBT administration</td>
<td>126.00</td>
<td>103.00</td>
<td>$229.00</td>
</tr>
<tr>
<td>Scoring</td>
<td>3.00</td>
<td>42.00</td>
<td>$45.00</td>
</tr>
<tr>
<td>Program management</td>
<td>7.00</td>
<td>7.00</td>
<td>$14.00</td>
</tr>
<tr>
<td>Educational Materials</td>
<td>4.00</td>
<td>4.00</td>
<td>$8.00</td>
</tr>
<tr>
<td>Overhead</td>
<td>14.00</td>
<td>17.00</td>
<td>$31.00</td>
</tr>
<tr>
<td><strong>Total Ideal Exam</strong></td>
<td><strong>$178.00</strong></td>
<td><strong>$198.00</strong></td>
<td><strong>$376.00</strong></td>
</tr>
</tbody>
</table>

(1) See supporting schedule and footnotes on Page 71
(2) See supporting schedule and footnotes on Page 72
<table>
<thead>
<tr>
<th>Task</th>
<th>Inflation Factor (1)</th>
<th>No. of Personnel</th>
<th>% of Time</th>
<th>Annual Salary (2)</th>
<th>Total</th>
<th>Per person Direct Cost Estimated Unit Costs</th>
<th>Grand Total</th>
<th>Cost per Candidate (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Per person</td>
<td>QTY.</td>
<td>UNIT.</td>
</tr>
<tr>
<td>Examination Development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item development (4)</td>
<td>17.0%</td>
<td></td>
<td>1</td>
<td>100%</td>
<td>96,000</td>
<td>112,320</td>
<td>2,000</td>
<td>849</td>
</tr>
<tr>
<td>Form Development (24 AM sets and 24 PM sets)</td>
<td>17.0%</td>
<td></td>
<td>1</td>
<td>100%</td>
<td>96,000</td>
<td>112,320</td>
<td>2,000</td>
<td>849</td>
</tr>
<tr>
<td>One Technical Manager</td>
<td>17.0%</td>
<td></td>
<td>2</td>
<td>100%</td>
<td>48,000</td>
<td>112,320</td>
<td>2,000</td>
<td>849</td>
</tr>
<tr>
<td>Two Technical Assistants</td>
<td>17.0%</td>
<td></td>
<td>2</td>
<td>100%</td>
<td>60,000</td>
<td>140,400</td>
<td>2,000</td>
<td>849</td>
</tr>
<tr>
<td>Form transfer and verification</td>
<td>17.0%</td>
<td></td>
<td>2</td>
<td>100%</td>
<td>60,000</td>
<td>140,400</td>
<td>2,000</td>
<td>849</td>
</tr>
<tr>
<td>Two editorial and production personnel</td>
<td>17.0%</td>
<td></td>
<td>2</td>
<td>100%</td>
<td>60,000</td>
<td>140,400</td>
<td>2,000</td>
<td>849</td>
</tr>
<tr>
<td>Sub-Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>365,040</td>
<td></td>
</tr>
<tr>
<td>CBT Vendor Administration 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eight hour test day</td>
<td>17.0%</td>
<td>1</td>
<td>1</td>
<td>100%</td>
<td>150,000</td>
<td>175,500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scoring (2003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentication and release to boards</td>
<td>17.0%</td>
<td></td>
<td>1</td>
<td>100%</td>
<td>90,000</td>
<td>105,300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interface with CBT vendor management of data transfer</td>
<td>17.0%</td>
<td></td>
<td>1</td>
<td>100%</td>
<td>17,550</td>
<td>175,500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Setting - Consultants</td>
<td>17.0%</td>
<td></td>
<td>1</td>
<td>15,000</td>
<td>17,550</td>
<td>175,500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>280,800</td>
<td></td>
</tr>
<tr>
<td>Program Management (2003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer services (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Customer Service Manager</td>
<td>17.0%</td>
<td>1</td>
<td>50%</td>
<td>54,000</td>
<td>31,590</td>
<td>31,590</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Six Customer Service Clerks</td>
<td>17.0%</td>
<td>6</td>
<td>50%</td>
<td>36,000</td>
<td>126,360</td>
<td>126,360</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project planning and management</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Executive Director</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Psychometrician</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two senior Technical Managers</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two Program Assistants</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel (Allocated half to each day) (7)</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>551,070</td>
<td></td>
</tr>
<tr>
<td>Educational Materials</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duplication of CD ROM</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revision of CD ROM</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other educational materials</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hardware Consultant</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Overhead (8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grand Total (9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(1) Inflation assumption: 4% per annum compounded, starting in the year 2000. Example: Salaries are estimated at January 1999 rates. 4% inflation is added for 2000, 8.2% for 2001, 12.4% for 2002, and 17% for 2003. (2) Salaries include estimated base salary plus 20% for benefits. (3) Based on 100,000 candidates testing. (4) Per item cost based on current AICPA cost for item development. (5) Based on mid-range estimates of current costs of $12 to $15 per hour ($13.50): adjusted by inflation rate of 4% per annum to calculate future costs. (6) 50% of effort is devoted to Day 1 and 50% to Day 2. (7) Includes travel costs for content area committees, Board of Examiners, COTF, Board visits and conferences, and Standard Setting Committee. (8) Overhead is applied to all costs except item development, which already has overhead built in. (9) Costs do not include NASBA fees.
<table>
<thead>
<tr>
<th>Task</th>
<th>Inflation Factor (1)</th>
<th>No. of Personnel</th>
<th>% of Time</th>
<th>Annual Salary (2)</th>
<th>Total</th>
<th>Direct Cost</th>
<th>Estimated Unit Costs</th>
<th>Grand Total</th>
<th>Cost per Candidate (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination Development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation Development [Takes place in 2003]</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form Development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Technical Manager</td>
<td>17.0%</td>
<td>1</td>
<td>100%</td>
<td>150,000</td>
<td>175,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two Technical Assistants</td>
<td>17.0%</td>
<td>2</td>
<td>100%</td>
<td>48,000</td>
<td>112,320</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Template Development</td>
<td>17.0%</td>
<td>1</td>
<td>100%</td>
<td>120,000</td>
<td>140,400</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Senior Technician</td>
<td>17.0%</td>
<td>1</td>
<td>100%</td>
<td>120,000</td>
<td>140,400</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form transfer and verification</td>
<td>17.0%</td>
<td>4</td>
<td>100%</td>
<td>60,000</td>
<td>280,800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>709,020</strong></td>
<td><strong>2,534,220</strong></td>
</tr>
<tr>
<td>CBT Vendor Administration (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Six and one half hour test day</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scoring</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free Response Scoring</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scoring - (method 1) 4 simulations (5)</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scoring - (method 2) 4 simulations (5)</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviewers' adjudication of discrepant scores (6)</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free response scoring - joining and verification</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program director</td>
<td>17.0%</td>
<td>1</td>
<td>100%</td>
<td>96,000</td>
<td>112,320</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical Programmer</td>
<td>17.0%</td>
<td>1</td>
<td>100%</td>
<td>90,000</td>
<td>105,300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation scoring, verification and joining</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program director</td>
<td>17.0%</td>
<td>1</td>
<td>100%</td>
<td>96,000</td>
<td>112,320</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical Programmer</td>
<td>17.0%</td>
<td>1</td>
<td>100%</td>
<td>90,000</td>
<td>105,300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Setting - Consultants</td>
<td>17.0%</td>
<td>1</td>
<td>100%</td>
<td>90,000</td>
<td>105,300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>435,240</strong></td>
<td><strong>3,761,550</strong></td>
</tr>
<tr>
<td>Program Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer services (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One customer service manager</td>
<td>17.0%</td>
<td>1</td>
<td>50%</td>
<td>54,000</td>
<td>31,590</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Six customer service clerks</td>
<td>17.0%</td>
<td>6</td>
<td>50%</td>
<td>36,000</td>
<td>126,360</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project planning and management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One executive director</td>
<td>17.0%</td>
<td>1</td>
<td>50%</td>
<td>180,000</td>
<td>105,300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Psychometricist</td>
<td>17.0%</td>
<td>1</td>
<td>50%</td>
<td>96,000</td>
<td>56,160</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two senior program directors</td>
<td>17.0%</td>
<td>2</td>
<td>50%</td>
<td>150,000</td>
<td>175,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two program assistants</td>
<td>17.0%</td>
<td>2</td>
<td>50%</td>
<td>48,000</td>
<td>56,160</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel (Allocated half to each day) (8)</td>
<td>17.0%</td>
<td>15,000</td>
<td></td>
<td></td>
<td>17,550</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>551,070</strong></td>
<td><strong>723,528</strong></td>
</tr>
<tr>
<td>Educational Materials</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duplication of CD ROM</td>
<td>17.0%</td>
<td>125,000</td>
<td>1,80</td>
<td>263,250</td>
<td>263,250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revision of CD ROM</td>
<td>17.0%</td>
<td>1</td>
<td>50,000</td>
<td>58,500</td>
<td>58,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other educational materials</td>
<td>17.0%</td>
<td>1</td>
<td>50,000</td>
<td>58,500</td>
<td>58,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hardware Consultant</td>
<td>17.0%</td>
<td>1</td>
<td>25,000</td>
<td>29,250</td>
<td>29,250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>409,500</strong></td>
<td><strong>409,500</strong></td>
</tr>
<tr>
<td>**Overhead (9)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>1,673,679</strong></td>
<td><strong>17.00</strong></td>
</tr>
<tr>
<td><strong>Grand Total (10)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>1,267,110</strong></td>
<td><strong>14,437,800</strong></td>
</tr>
</tbody>
</table>

(1) Inflation assumption: 4% per annum compounded starting in the year 2000.
(2) Salaries include estimated base salary plus 20% for benefits
(3) Based on 100,000 candidates testing
(4) Based on mid-range estimates of current costs of $12 to $15 per hour ($13.50): adjusted by inflation rate of 4% per annum to calculate future costs.
(5) Method 1 utilizes manual scoring, method 2 relies on automated scoring process
(6) Based on AICPA current cost for essay scoring
(7) 50% of effort is devoted to Day 1 and 50% to Day 2
(8) Includes travel costs for content area committees, Board of Examiners, COTF, Board visits and conferences, and Standard Setting Committee
(9) Overhead is applied to all costs except item development, which already has overhead built in.
(10) Costs do not include NASBA fees
### Table 3 - Implementation of Linear Exam

Estimated on a per-candidate basis for 100,000 candidates

<table>
<thead>
<tr>
<th></th>
<th>7 hr. Exam (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination development</td>
<td>$43.00</td>
</tr>
<tr>
<td>CBT administration</td>
<td>111.00</td>
</tr>
<tr>
<td>Scoring</td>
<td>3.00</td>
</tr>
<tr>
<td>Program management</td>
<td>9.00</td>
</tr>
<tr>
<td>Overhead</td>
<td>13.00</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>$179.00</strong></td>
</tr>
</tbody>
</table>

(1) See supporting schedule and footnotes on Page 74
### Table 3 Supporting Schedule

#### Linear Examination - Implementation - Seven Hour Exam

<table>
<thead>
<tr>
<th>Task</th>
<th>Inflation Factor (1)</th>
<th>No. of Personnel</th>
<th>% of Time</th>
<th>Annual Salary (2)</th>
<th>1999 Total</th>
<th>1999 Estimated Unit Cost</th>
<th>1999 Grand Total</th>
<th>( Rounded ) Cost per Candidate (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form transfer and verification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two editorial and production personnel</td>
<td>17.0%</td>
<td>2</td>
<td>120.0%</td>
<td>60,000</td>
<td>168,480</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical manager</td>
<td>17.0%</td>
<td>1</td>
<td>120.0%</td>
<td>96,000</td>
<td>134,784</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical Assistant</td>
<td>17.0%</td>
<td>1</td>
<td>120.0%</td>
<td>48,000</td>
<td>67,392</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>202,176</strong></td>
<td></td>
<td><strong>4,276,056</strong></td>
</tr>
<tr>
<td>CBT vendor administration (3,4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seven hour test day</td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>700,000</strong></td>
<td><strong>11,056,500</strong></td>
<td><strong>11,056,500</strong></td>
</tr>
<tr>
<td>Scoring</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentication and release to boards</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program Director</td>
<td>17.0%</td>
<td>1</td>
<td>100.0%</td>
<td>150,000</td>
<td>175,500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interface with CBT vendor management of data transfer</td>
<td>17.0%</td>
<td>1</td>
<td>100.0%</td>
<td>90,000</td>
<td>105,300</td>
<td>35,100</td>
<td>105,300</td>
<td>3.00</td>
</tr>
<tr>
<td>Statistical Programmer</td>
<td>17.0%</td>
<td>1</td>
<td>100.0%</td>
<td>90,000</td>
<td>105,300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>280,800</strong></td>
<td></td>
<td><strong>315,900</strong></td>
</tr>
<tr>
<td>Program Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer services</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Customer Service Manager</td>
<td>17.0%</td>
<td>1</td>
<td>50.0%</td>
<td>54,000</td>
<td>31,590</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Six Customer Service Clerks</td>
<td>17.0%</td>
<td>6</td>
<td>50.0%</td>
<td>36,000</td>
<td>126,360</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project planning and management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One executive director</td>
<td>17.0%</td>
<td>1</td>
<td>50.0%</td>
<td>180,000</td>
<td>105,300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One Psychometrician</td>
<td>17.0%</td>
<td>1</td>
<td>50.0%</td>
<td>96,000</td>
<td>56,160</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two senior technicians</td>
<td>17.0%</td>
<td>2</td>
<td>50.0%</td>
<td>150,000</td>
<td>175,500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two technical assistants</td>
<td>17.0%</td>
<td>2</td>
<td>50.0%</td>
<td>36,000</td>
<td>42,120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Travel (5)</strong></td>
<td>17.0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>344,916</strong></td>
<td></td>
<td><strong>344,916</strong></td>
</tr>
<tr>
<td><strong>Sub-Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>537,830</strong></td>
<td></td>
<td><strong>881,246</strong></td>
</tr>
<tr>
<td>Overhead (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>1,262,500</strong></td>
</tr>
</tbody>
</table>

(1) Inflation assumption: 4% per annum compounded starting in the year 2000.
(2) Salary costs include 20% additional for fringe benefits
(3) Based on 100,000 candidates testing
(4) Based on mid-range estimates of current costs of $12 to $15 per hour ($13.50). adjusted by inflation rate of 4% per annum to calculate future costs.
(5) Includes travel costs for content area committees, Board of Examiners, COTF, Board visits and conferences, and Standard Setting Committee
(6) 10% overhead is added to all costs except item development