1. Letter of Submittal
November 7, 2011

Michael Middleton
Director, Assessment Business and Special Populations
Assessment and Student Information
Office of Superintendent of Public Instruction
600 Washington Street South
Olympia, WA 98501

Dear Mr. Middleton:

Educational Testing Service (ETS) is pleased to submit our proposal in response to the SMARTER Balanced Assessment Consortium’s RFP No. 05 for Psychometric Services. We are eager to collaborate with member states of the Consortium and your Technical Advisory Committee to design and provide the highest quality psychometric services for the summative and interim components of the assessment system as it evolves during the next three years.

We share the Consortium’s goal of providing technically sound, reliable, valid, and educationally meaningful assessment results. ETS has a large staff of qualified experts and deep experience in the type of psychometric design and service called for in the RFP. Along with our collaborative partner for this effort, Measured Progress, we offer our best and brightest team to SBAC to work together to meet the challenging technical tasks required.

ETS and Measured Progress are mission driven not-for-profit companies and have each built a reputation for integrity in our psychometric and research work. We appreciate the ground-breaking nature of your Race to the Top assessment program and will be honored to support SBAC in the development of a system that is innovative, teacher- and student-oriented, and sustainable.

Please let me know if you have any questions.

Sincerely,

John H. Oswald
Vice President and General Manager

www.ets.org
October 24, 2011

John Oswald  
Vice President and General Manager,  
ETS Multistate Assessment Programs  
Educational Testing Service  
Corporate Headquarters – Mail Stop 58L  
Rosedale Road  
Princeton, NJ 08541

Dear Mr. Oswald:

Measured Progress is very pleased to offer its support, as a subcontractor to ETS, in its bid for the SMARTER Balanced Assessment Consortium (SBAC) Psychometric Services contract in response to SBAC RFP No. 05. Should ETS be selected for contract award, Measured Progress will fulfill its obligations to ETS as outlined in the scope of work described in the proposal.

We are pleased to provide personnel, technical services, and expertise to perform specific tasks of this scope of work related to the following: standard-setting design, psychometric consultation, and participation in TAC meetings.

Measured Progress welcomes the opportunity to work on this critically important program.

Regards,

Martin S. Borg  
President and CEO
TABLE OF CONTENTS

1. Letter of Submittal .................................................................................................................................. 1-1
   Exhibit A – Certifications and Assurances

2. Technical Proposal .................................................................................................................................... 2-1
   Executive Summary ................................................................................................................................. 2-1
   Overview of Our Proposed Scope of Work ............................................................................................... 2-4
   A. Project Approach/Methodology ........................................................................................................ 2-13
      2.1 Determine Linking and Equating Design for Special Forms ....................................................... 2-13
      2.2 Develop Technical Manuals ....................................................................................................... 2-28
      2.3 Determine Standard-Setting Design .......................................................................................... 2-34
      2.4 Determine Vertical Scale Design ............................................................................................... 2-41
      2.5 Design Pilot Test Sampling Plan and Select Items and Tasks for Pilot Test Forms ................. 2-56
      2.6 Develop Pilot Test Item and Task Data Review Materials ......................................................... 2-61
      2.7 Conduct Psychometric Analysis to Support Pilot Test Data Review ..................................... 2-65
      2.8 Conduct Analysis of Pilot Test ................................................................................................... 2-88
      2.9 Design Item and Student Sampling Plans for Field Test and Select Anchor Items and
      Tasks for Calibration and Building the Vertical Scale ................................................................. 2-90
      2.10 Final Field Test Forms Verification ....................................................................................... 2-98
      2.11 Conduct Psychometric Analysis to Support Field Test Data Review ................................... 2-104
      2.12 Conduct Psychometric Analysis to Support Item and Task Calibration ............................... 2-107
      2.13 Present to TAC meetings as required ................................................................................... 2-109
   B. Work Plan ......................................................................................................................................... 2-112
   C. Project Schedule ............................................................................................................................... 2-122
   D. Deliverables ....................................................................................................................................... 2-124
   E. Outcomes and Performance Measurement .................................................................................. 2-129
   F. Risks .................................................................................................................................................. 2-136

3. Management Proposal .......................................................................................................................... 3-1
   A. Project Management ....................................................................................................................... 3-1
   B. Experience of the Consultant ........................................................................................................ 3-28
   C. References ....................................................................................................................................... 3-98
   D. Contractor Intake Form (Exhibit E) ............................................................................................... 3-109

4. Cost Proposal ......................................................................................................................................... 4-1

Appendix
   Reference Citations
   ETS Résumés
   Measured Progress Résumés
   External Consultant Résumés

SBAC RFP No. 05
NOTICES

Certain information and data contained herein are proprietary and are marked as confidential in accordance with the RFP specifications. This information and data shall not be duplicated, used, or disclosed other than to evaluate this proposal for possible funding; provided that, if a contract is awarded to Educational Testing Service as a result of the submission of this proposal, SBAC shall have the right to duplicate, use or disclose the information or data to the extent provided in the RFP specifications in conformance with federal law. ETS restricts the release of its price proposal to the extent provided by the RFP and the applicable FOIA.

Any request for release of information or data under any freedom of information request will be for the use of the individual requestor only, and will not constitute a license to publish or distribute, or a transfer of any rights in or to the information or data. Paragraphs and/or pages stamped “confidential” shall be redacted from the final released version of the ETS proposal.

Copyright © 2011 Educational Testing Service

Copyright in this material is owned by Educational Testing Service, and limited dissemination has been made for the above purpose only. Except as provided above, no reproduction, use or distribution of any part of this material may be made without the specific authorization of Educational Testing Service.
ETS Letter of Submittal Attachment

1. Name, address, principal place of business, telephone number, fax number and email address of legal entity or individual with whom contract would be written:

| Contact: | John Oswald (For signature purposes)  
Vice President & General Manager, Multistate Assessment Programs  
10999 Interstate Highway 10 West, Suite 400, San Antonio, TX 78231  
Phone: (210) 558-5633 / Fax: (210) 558-5636  
Email: joswald@ets.org |
| Contract Representative: | Jean Shipos, Contract Manager  
Educational Testing Service  
Mail Stop 58L, Rosedale Road, Princeton, NJ 08541  
Phone: 609-683-2735 / Fax: 609-683-2766  
Email: jshipos@ets.org |
| Corporate Headquarters: | Educational Testing Service (ETS)  
Corporate Contracts, Mail Stop 58L  
Rosedale Road, Princeton, New Jersey 08541 |

2. The name of the contact person for this RFP
   John Oswald, Vice President & General Manager, Multistate Assessment Programs  
   Phone: (210) 558-5633 / Fax: (210) 558-5636  
   Email: joswald@ets.org

3. Name, address, and telephone number of each principal officer (President, Vice President, Treasurer, Chairperson of the Board of Directors, etc.)

| Kurt Landgraf, President & CEO | Walt MacDonald, Executive VP & COO Programs, Services & Operations |
| Ida Lawrence, SVP Research & Development | Stephen Lazer, VP Student & Teacher Assessments |
| Scott Paris, VP Research | John Oswald, VP & GM Multistate Assessment Programs |
| John Mazzeo, VP Statistical Analysis & Psychometric Research | Glenn Schroeder, Senior VP & General Counsel |
| Marisa Farnum, VP Assessment Development | Jack Hayon, Senior VP & CFO, Finance & Administration |

All officers can be reached at:
Educational Testing Service Corporate Headquarters  
Mail Stop 58L, Rosedale Road, Princeton, NJ 08541  
609-683-2735

4. Legal status of the Consultant (sole proprietorship, partnership, corporation, etc.) and the year the entity was organized to do business as the entity now substantially exists

Educational Testing Service (ETS) is a non-stock, non-profit corporation focused on assessment development and educational research. ETS is organized and existing under the Education Law of the State of New York. ETS was incorporated in 1947.

5. Location of the facility from which the Consultant would operate
   K-12 Assessment Programs Division:  
   10999 Interstate Highway 10 West, Suite 400, San Antonio, TX 78230  
   ETS Headquarters: Rosedale Road, Princeton, NJ 08541

6. A detailed list of all materials and enclosures included in the Proposal
ETS submits the following with our proposal response:
   » ETS Technical Proposal (including Appendix)
   » ETS Cost Proposal
Measured Progress Letter of Submittal

1. Name, address, principal place of business, telephone number, fax number and email address of legal entity or individual with whom contract would be written:

   Measured Progress
   100 Education Way, Dover, NH 03820
   Phone: 603-749-9102 / Fax: 603-749-6398
   Email: borg.martin@measuredprogress.org

2. The name of the contact person for this RFP
   Michael L. Nering, Ph.D., Assistant Vice President | Research & Analysis
   Phone: 603.749.9102, ext. 2192
   Email: Nering.Michael@measuredprogress.org

3. Name, address, and telephone number of each principal officer (President, Vice President, Treasurer, Chairperson of the Board of Directors, etc.)

   Martin Borg, President and Chief Executive Officer
   Richard Dobbs, Senior Vice President
   Timothy Crockett, Senior Vice President
   Richard Swartz, Senior Vice President
   Lisa Ehrlich, Chief Operating Officer
   Thomas Squeo, Chief Information Officer
   Robert Kmetz, Chief Financial Officer and Treasurer
   Mike Russell, Vice President of Nimble Innovation Lab
   Dirk Mattson, Vice President Assessment Design, Development and Operations

   All can be reached at:
   Measured Progress
   100 Education Way, Dover, NH 03820
   Phone: 603-749-9102

   Board Chair:
   Alice J. Irby
   680 Lake Forest Drive SE, Pinehurst, NC 28374
   Phone: 910-215-4584

4. Legal status of the Consultant (sole proprietorship, partnership, corporation, etc.) and the year the entity was organized to do business as the entity now substantially exists
   Measured Progress is a 501(c) 3 not-for-profit corporation. It was founded in 1985.

5. Location of the facility from which the Consultant would operate
   100 Education Way
   Dover, NH 03820
EXHIBIT A
CERTIFICATION AND ASSURANCES

I/we make the following certifications and assurances as a required element of the proposal to which it is attached, understanding that the truthfulness of the facts affirmed here and the continuing compliance with these requirements are conditions precedent to the award or continuation of the related contract(s):

1. I/we declare that all answers and statements made in the proposal are true and correct.

2. The prices and/or cost data have been determined independently, without consultation, communication, or agreement with others for the purpose of restricting competition. However, I/we may freely join with other persons or organizations for the purpose of presenting a single proposal.

3. The attached proposal is a firm offer for a period of 60 days following receipt, and it may be accepted by OSPI without further negotiation (except where obviously required by lack of certainty in key terms) at any time within the 60-day period.

4. In preparing this proposal, I/we have not been assisted by any current or former employee of the state of Washington whose duties relate (or did relate) to this proposal or prospective contract, and who was assisting in other than his or her official, public capacity. (Any exceptions to these assurances are described in full detail on a separate page and attached to this document.)

5. I/we understand that OSPI will not reimburse me/us for any costs incurred in the preparation of this proposal. All proposals become the property of OSPI, and I/we claim no proprietary right to the ideas, writings, items, or samples, unless so stated in this proposal.

6. Unless otherwise required by law, the prices and/or cost data which have been submitted have not been knowingly disclosed by the Vendor and will not knowingly be disclosed by him/her prior to opening, directly or indirectly to any other Vendor or to any competitor.

7. I/we agree that submission of the attached proposal constitutes acceptance of the solicitation contents and the attached sample contract and general terms and conditions. If there are any exceptions to these terms, I/we have described those exceptions in detail on a page attached to this document.

8. No attempt has been made or will be made by the Vendor to induce any other person or firm to submit or not to submit a proposal for the purpose of restricting competition.

9. I/we grant OSPI the right to contact references and others, who may have pertinent information regarding the Vendor’s prior experience and ability to perform the services contemplated in this procurement.

On behalf of the firm submitting this proposal, my name below attests to the accuracy of the above statements.

Signature of Vendor

Vice President and General Manager

Title

Date

11/4/2011

Page 31 of 51  RFP No. 05: Psychometric Services
ETS Exceptions for RFP-05

The RFP gives the potential Vendor the opportunity to submit exceptions to the General Contract Terms & Conditions (Paragraph 11, page 13 of the RFP). In reviewing the standard template (Attachment A) that will become the basis for the RFP-05 contract award, ETS respectfully requests that the following changes be considered:


   Explanation:
   The RFP requires that the Contractor use third party, commercials, or Open Source computer programs and methodologies. This impacts the contract template and the right(s) of the Contractor to grant unrestricted use to such computer programs and/or methodologies. Please add the highlighted phrase for clarification regarding Superintendent’s access to third party and/or Open Source computer models and methodology.

   Revision:
   “In compliance with Chapter 39.29 RCW, the Contractor shall provide access to data generated Under this contract to the Superintendent, the Joint Legislative Audit and Review Committee and the State Auditor at no additional cost. This includes access to all information that supports the findings, conclusions, and recommendations of the Contractor’s reports, including computer models and methodology for those models, providing that ETS has the right to provide access to third party computer models and/or methodologies. If such Materials are not owned by ETS, the Superintendent may have to register for the appropriate license."


   Explanation:
   Please revise the standard contract language, using the highlighted phrases below in order to clarify the ability of the Contractor to grant unlimited rights to third party materials, including access to third party and/or Open Source computer models and methodology. ETS cannot guarantee unlimited rights and ownership, including patent rights, to Materials (as defined within Attachment A), if third party computer models and methodologies are used.

   Revision:

   (a) Second paragraph – Please delete the highlighted phrases: “Materials means all items in any format and includes, but is not limited to, data, reports, documents, pamphlets, advertisements, books, magazines, surveys, studies, computer programs, films, tapes, and/or sound reproductions. Ownership includes the right to copyright, patent, register and the ability to transfer these rights.”

   (b) Third paragraph – Please delete the paragraph in the contract in its entirely and replace with the following. “For Materials that are delivered under the contract, but that incorporate pre-existing materials not produced under the contract, the Contractor grants to the Agency the appropriate rights of use as defined by Contractor’s third party license(s) and/or Open Source permissions. For Materials that the Contractor pre-owns, or previously-developed, the Contractor grants more extensive rights to be defined based upon Superintendent’s needs and on the specific Material.”

Explanation:
The Contractor cannot assume responsibility for errors or delays created by the State. Significant penalties, based on Contractor’s delays and/or errors, are specified in the RFP. In order to protect itself, Contractor expects that the State will assume liability for errors or delays that the State (or other contracting party) creates/causes.

Revision:

Paragraph 2 - Please delete in its entirely and replace with the following.

“Contractor expressly agrees to indemnify, defend and hold harmless the Superintendent for any claim that is the direct result of Contractor’s or subcontractor’s performance or failure to perform under the terms of the contract. Contractor and/or Subcontractor, are not responsible for and shall not be financially penalized for delays in service or performance, or for errors in service or performance, which are directly attributable to Superintendent’s (or its agents’, employees’, or officials’) negligence or errors.”
2. Technical Proposal
Section 2 – Technical Proposal

Executive Summary

Educational Testing Service (ETS) is very pleased to respond to the SMARTER Balanced Assessment Consortium’s RFP No. 05 to provide psychometric services to the Consortium. As in other proposals to the SBAC, ETS is collaborating with Measured Progress on this submission. We would be honored to be chosen as the psychometric services partner for the Consortium.

This Executive Summary has two sections. First is a description of why the ETS/Measured Progress Collaborative is ideally suited to provide the psychometric services required in the SBAC RFP. Second is a high-level overview of our proposed work plan.

Why ETS and Measured Progress

When U.S. Secretary of Education Arne Duncan announced on April 6, 2010, the establishment of a $350 million grant under the Race to the Top program for the creation of comprehensive, next generation assessment systems, the testing industry certainly took notice. We knew that this was going to change our world in significant ways. With few small exceptions, such as the New England Common Assessment Program (NECAP) and the American Diploma Project’s Algebra Assessment, K–12 standardized assessment systems had always been designed and delivered either for individual states, school districts, or the U.S. Department of Education. During the ensuing months, we all — states, testing providers, and other organizations supporting assessment — asked many questions: Who would do the work? How would it be broken up? How would the procurement take place?

As the testing companies have tried to think through their futures and their roles in this historic effort, the states themselves have been going through even more careful examination, and in many cases, uncomfortable change. State standards, in most cases developed over years of debate and compromise, are now subject to replacement by the Common Core State Standards (CCSS). State assessment systems are being questioned and examined for replacement. While some people may feel that the new standards and proposed assessments are better than the existing ones, the existing systems may feel more comfortable. District test directors, principals, and teachers understand the existing systems. The state education agencies have, over many years and with public communication, familiarized the press and the public with what the state assessment is and how it works. Now all of these things are subject to change.

Most importantly, states now have had to learn to work together for a common goal. This requires compromise, and often it seems that the state’s delegate to the consortium’s deliberations, usually the state test director, has had less than absolute support for these efforts back home. Governors and legislators, unions, and associations of educators in the state have
differing opinions as to the wisdom of participation. Not only does the state test director have to negotiate his or her state’s needs with the rest of the consortium’s members, but they also have to negotiate with their own departments, boards, and governments in their state capitol. In other words, the members of the SBAC may be functioning outside their comfort zones for more than a year now. The driving force that makes this discomfort (and extremely hard work) worth it is the vision of a next generation assessment system that will be significantly better than what we have had in the past. Of equal importance is a theory of action that shows we really have a chance at achieving the goal of having every high school graduate college and career ready. As we have interacted with the SBAC members and leaders over the past year at those venues created for our discussions together, ETS and Measured Progress realized that we share your goals. That means that we also should be willing to share your commitment and should be willing to work outside of our comfort zones.

The nature of the conversation between ETS and Measured Progress as we considered working together, and hopefully securing an opportunity to work with the SBAC on delivering the work of the program, has been dramatically different from what one might expect. It has been much less about our business needs and much more about creating a working relationship across multiple levels of our organizations that would lead to the success of the SBAC and to your goals being achieved. This has involved a different sort of conversation in which the questions really became, “What can we do together that we could not do separately to help the SBAC be successful?”

The first step involved our creating a teaming agreement that would, most importantly, facilitate the creation of a teaming environment. We agreed that in order to achieve your objectives, we would need to work together as true partners, collaborating with each other, and very likely with others, in order to be successful. We agreed that one of our two organizations would have to serve as a prime contractor and one as a subcontractor in our submissions, and sometimes that choice would be different for each RFP. For example, we bid on SBAC-04 and SBAC-06 with Measured Progress as the prime contractor and ETS as the subcontractor. On this procurement, SBAC-05, we decided it would better suit the needs of the Consortium for ETS to be the prime contractor.

Regardless of the prime/sub relationship, we agreed to look to our mutual strengths in establishing teams and leaders that brought the best talent that we collectively have to offer. As a result, our proposal includes both companies in leading aspects of the project. Although each organization brings considerable psychometric expertise to this effort, we put together a structure that allows both of us to work in all areas, while emphasizing the areas in which one company might bring a particular advantage. By having each company lead to its strength, we are able to provide the SBAC with a best-in-class solution. ETS has an unparalleled history of research in the field of assessment and is known for attracting and retaining talented measurement professionals, producing high-quality assessments, and being willing and able to meet new industry challenges. Measured Progress is known for quality, a superior client service
orientation, innovation in areas such as accessibility and technology-based assessment, and flexibility. Both organizations have established credibility in K–12 educational assessments and have vast experience in building standards-based assessments. Both have worked extensively in all aspects of creating customized assessments and have worked successfully with many states that are members of the SBAC.

Measured Progress is already supporting ETS in one state program, and we are working side by side as prime contractors in another. Both of us are also very accustomed to collaborating with other organizations. For example, at ETS we work with other contractors in every one of our nine state contracts, sometimes as a prime contractor, and sometimes as a subcontractor. We believe that the challenging goals driving this project are best achieved by our working closely together in true collaboration.

Both ETS and Measured Progress are not-for-profit companies. The significance of this to the SBAC is that we have no shareholders to make wealthy, so our only goal is to help our clients be successful. Furthermore, neither of our companies has any product to sell, other than the services we provide to our clients. We will not be concerned with finding a way to convince the SBAC that some existing product we've developed is a better solution than the custom-made one you seek through this RFP and in your Application. Through our collective references and the structure of the work process outlined in our proposal, we hope it is evident that we take very seriously the inclusion of client stakeholders and the opportunity they have to influence the direction we take. Our well-thought-out, structured processes will allow flexibility while at the same time applying rigor, organization, and effective project management. We support our clients, providing frequent communication, transparency (no surprises), access to senior management, and the right knowledgeable and dedicated people on the project.

We believe that together we have the depth of experience and capacity to support the achievement of the SBAC’s primary goal: the development and delivery of a sustainable assessment system that supports the Theory of Action expressed in the SBAC application (Washington State, 2010). What follows is our plan for providing psychometric services for the SMARTER Balanced Assessment Consortium.
Overview of Our Proposed Scope of Work

2.1 Determine Linking and Equating Design for Special Needs Forms

Introduction

Given the innovative nature of the proposed assessment system, and the stakes that will be associated with results, the Consortium is seeking a pilot test designed to address issues/questions well in advance of operational implementation. The pilot design poses a considerable challenge given the wide variety of purposes the data collected are to serve. These purposes include evaluating the performance characteristics of summative items and performance tasks, calibrating and linking items and tasks both horizontally and vertically to defined proficiency scales, examining the impact of test dimensionality on scores and proficiency scales, and evaluating the performance and comparability of both paper and accommodated test forms. The diversity of these requirements demands a design that is necessarily complicated in its details. It is therefore best to provide an overview of the general design here before each of its components is discussed in detail by later sections of this proposal. For example, while this section will cover the general design, detailed plans for establishing the vertical scale are covered in Section 2.4. Similarly, pilot anchor and form construction is covered in Section 2.5, and the analyses completed following the data collection are covered in Sections 2.7 and 2.8.

We propose a matrix data collection design that uses both common-item and equivalent groups links to bind all items and tasks at a given grade level to a common scale. The design also incorporates common-item links between grade levels to establish a vertical scale. These links are implemented by administering at most grade levels blocks of test content sampled from the immediately lower grade level. Administering out-of-level content at lower but not higher grade levels minimizes concerns regarding opportunity to learn.

The requirement in our first proposed design is that all students are administered both a form of items and a performance task component. While this is an advantage in terms of strength of the linking, it can also be a disadvantage in terms of burden to teachers and students. If the testing time per student is impractical, then we offer a second design as an alternative. Under this design, most students are administered either a form or a performance task component.

The goal for the pilot is to collect approximately 1,500 responses per item and task. This is required to support the investigation of the various IRT models as well as for the dimensionality studies. Such sample sizes are appropriate with new and innovative testing programs, and they permit obtaining calibrations for items and tasks with a range of difficulties and evaluating differential item functioning (DIF) for a variety of groups of students. For the paper and pencil comparability studies, a subset of schools would be chosen to identify two classrooms per
tested grade: one for online administration and one for paper administration. It would be ideal if students from these classrooms could then be randomly assigned to testing mode. An alternative would be to identify matching schools and assign one from each pair to each method.

Careful evaluation of equating results is essential prior to, during, and after IRT scaling to examine the integrity of the vertical scales, as well as the quality of the equating of forms within grade. To evaluate the horizontal equating, we will also examine various classical test theory statistics, as well as those listed as part of the IRT analyses.

The pilot test design also allows for the evaluation of the performance tasks and associated scoring processes. We propose to conduct a set of psychometric studies to evaluate the performance tasks, which will be used for the data review meetings described in Section 2.6 to evaluate the items and make recommendations for writing improvements. These statistics will also be uploaded to the Consortium item bank. All of the data collection designs described in this proposal will yield sufficient numbers of examinees complete with the classical item analysis, differential item functioning (DIF), item response theory (IRT) analysis, rater analysis, trend studies and generalizability analysis, which are described in detail in Section 2.7.

In order to place the paper and pencil and the accommodated or modified forms onto the vertical scale established for the adaptive test, a special data collection and study will be completed during the pilot test. We propose to use online forms for the paper and pencil and special needs forms, with the goal of having the majority of the items administered in both online and paper and pencil modes. This will allow for a common-item linking design to place the fixed forms on the vertical scale and will provide the data for the comparability study described in Section 2.7. We will also provide the required written descriptions of all planned activities in a timely fashion for SBAC review and approval, and will conduct the twice-annual contractor meetings required.

2.2 Deliver Technical Manuals

We will deliver an on-time and accurate SBAC Technical Manual. We are dedicated to working with the Consortium and its Technical Advisory Committee to design comprehensive documentation that will provide evidence to support the validity of all aspects of the SBAC pilot and field tests. This document will serve as the ultimate source of information about the SBAC assessment program and will be designed to serve the needs of various educators and stakeholders of the Consortium. For that reason, we are envisioning one master document that is essentially a compilation of various stand-alone technical reports. That way, constituents can refer to and reference a certain chapter of the Technical Manual without having to read the entire document or refer to other sections for context. Of course, the primary goal of the Technical Manual is to meet the needs of the Consortium, and as such, we will collaborate to
design the Technical Manual(s) to the exact specifications desired by the Consortium and its technical advisors.

The Consortium may also be interested in developing a technical documentation website, with hyperlinks to the Technical Manuals and related documentation, as well as corresponding data access. Should that be of interest, we will draw on our experience with the NAEP technical documentation website. We also understand the considerable planning and effort that will be required to assemble information from the various Contractors to help ensure that all aspects of the SBAC system are completely transparent. We will draw upon our direct experience in working successfully with a number of Contractors to create technical manuals for our current clients.

2.3 Determine Standard-setting Design

Together, ETS and Measured Progress provide a vast array of experience and expertise in standard setting. ETS’s collaborator and subcontractor, Measured Progress, will have the lead role in determining the standard-setting design. Measured Progress has extensive experience in conducting standard setting meetings. Results of their work have been presented at various national professional meetings, and were included in a chapter entitled “Setting Performance Standards Using the Body of Work (BoW) Method” in Setting Performance Standards: Concepts, Methods, and Perspectives, edited by Gregory J. Cizek and released in 2001. Measured Progress has implemented standard-setting sessions in many states including Kentucky, Maine, Massachusetts, Montana, New Hampshire, Rhode Island, Vermont, Florida, Missouri, Nevada, New York, New Jersey, Utah, South Carolina, and Colorado. Additionally, Measured Progress is currently working with the National Assessment Governing Board (NAGB) to assist with setting standards for college and career readiness in Reading and Mathematics for the National Assessment of Educational Progress (NAEP). Their work on this contract has allowed them to develop a paperless form of standard setting where all presented materials and data collection are done using an online process. Measured Progress’s involvement in this project will position us to help SBAC establish its definitions and performance standards for college and career readiness.

ETS is a leading measurement organization in large-scale standard setting. ETS’s Research and Development Division includes highly regarded and established thought leaders, researchers, and practitioners in the area of standard setting. Recent contributions to the field of standard setting include Bejar, Braun, and Tannenbaum’s chapter in Assessing and Modeling Cognitive Development in School (2007); Hambleton and Pitoniak’s chapter in the latest edition of Educational Measurement (2006); Tannenbaum and Katz’s chapter in the forthcoming first edition of the APA Handbook of Testing and Assessment in Psychology (in press); and Zieky, Perie, and Livingston’s book Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests (ETS, 2008). ETS has experience working with cross-state standard setting through its development and implementation of a multistate standard-setting
process for educator licensure tests (Tannenbaum, 2011), and has previous experience conducting standard setting for computer adaptive tests (O’Neill, Tannenbaum, & Tiffen, 2005).

Our process will be designed to not only draw on the collective experience of the ETS/Measured Progress partnership, but will include an array of expertise from the larger measurement community. Together, ETS, Measured Progress, the SBAC, and the external consultants will form the standard-setting design committee (SSDC) for the SBAC. Members of the SSDC will be recognized leaders in the field of psychometrics, and they will have significant expertise in the area of standard setting. SSDC members will be determined upon award of the contract.

2.4 Determine Vertical Scale Design

As the RFP indicates, until information is available regarding the dimensionality of each of the content area assessments, we feel it is unwise at this time to propose specific scaling approaches. We will work in collaboration with the SBAC and your TAC to specify the number and types of scales to develop in order to best reflect the underlying structure of the constructs and simultaneously provide scores that are both understandable and useful to the Consortium and your stakeholders. We also expect to involve the TAC in design decisions and interpretation of results periodically, as determined by the Consortium. In Section 2.4 of this Proposal, we describe our approach to investigating the impact of the construct dimensionality and our approach to constructing and evaluating a provisional scale with the pilot test data. We also include our proposal for establishing and evaluating the materials and procedures for scoring of the performance tasks.

2.5 Design Pilot Test Sampling Plan and Select Items and Tasks for Pilot Test Forms

To begin item selection for the pilot test, we will study the Consortium’s test specifications and blueprints for each content and grade to fully understand the expectations for how the developmental progression toward college and career readiness articulated in the Common Core State Standards (CCSS) will be operationalized in summative assessments and performance tasks for each grade level. Based on this review and in collaboration with the Consortium, its technical advisors, and content experts, we will develop specifications for the vertical linking items, and for the selection of on-grade items to populate the pilot test forms. Preliminary forms assembly guidelines will be developed for item selection to fill the requirements of the test blueprints and specifications, including those of the selected linking design. This is an iterative process in collaboration with all parties identified by the Consortium to finalize the test assembly guidelines.

This section of our proposal describes the sampling plan and item selection in great detail, including selection of anchor items and the pilot testing plan for the performance items.
2.6 Develop Pilot Test Item and Task Data Review Materials

After pilot testing, it will be important to involve educators from across the Consortium in a thorough review of all pilot test items and performance tasks, including information on the performance of items under different conditions (e.g., paper and pencil and translated versions). An examination of the actual statistical results from the pilot test for all items and performance tasks will be an important part of this review. Clear communication and presentation of information is paramount to a successful data review meeting, but will be even more critical given the introduction of new item types, new delivery methods, and content associated with the new CCSS. The data and information must be presented in a clear and concise manner to maximize the opportunity for educators to discuss the observed results and provide feedback that can be used to prepare items for field testing.

In the case of the performance tasks, there will need to be additional training for the educators attending the meetings on evaluating the data from these items and the scoring methods and systems, as this information may be unfamiliar to some attendees. As noted in the Q&A, due to the large volume of items, the data review meetings may be conducted online, through a distributed system, with face-to-face meetings limited, possibly to training or general orientation. ETS and Measured Progress have extensive experience in conducting both face-to-face and virtual meetings, and look forward to developing a system for data review that will meet the Consortium’s needs for educator engagement and efficiency in the review process. Our approach to preparing for these reviews is described below, and we expect to work closely with the Executive Director and Lead Psychometrician of the Consortium in planning and designing these meetings to maximize our opportunity to share and gather information with educators.

2.7 Conduct Psychometric Analysis to Support Pilot Test Data Review

The data collection design for the pilot test that we proposed and described in Section 2.1 will support several studies including:

1. Evaluation of item and performance tasks
2. IRT model selection
3. Test form information
4. Evaluation and monitoring of performance task scoring
5. Accessibility for students with disabilities and English learners
6. Comparability of paper and pencil and online versions

In Section 2.7, we describe in detail our proposed plan to complete each of these studies, along with the available methods and software that are available to use. The detail provided in Section 2.7 reflects the software that ETS believes could be useful to the Consortium, including our evaluation of how effective they are likely to be for Consortium work. Though we recommend
those with an overall rank of one as being likeliest to support consistent, accurate results, we are willing to work with the Consortium to create an effective solution with any of the software packages listed. ETS will make available any software code associated with any program as part of the technical documentation, as noted in Section 2.2. The results of these analyses will be provided to the Consortium, used in the pilot data review meeting, summarized and described in the Technical Manuals, and, where appropriate, loaded into the Consortium item bank.

2.8 Conduct Analysis of Pilot Test

After administration of the pilot test, we will work cooperatively with the assessment development Contractor and the Consortium to determine if any changes need to be made to the item development process. Some areas that we will examine include the implementation of innovative items, the performance of students in each standard, and the distribution of score points for each performance task. The use of innovative items especially will require a careful examination of how students interacted with the technology associated with the delivery system. If our analysis determines that improvements could be made in any area of item development, then we will work closely with the Consortium and the item development Contractor to identify and make any potential improvements in the process. Because a vertical scaling was developed, some plan for item usage and performance across grade levels might also be evaluated.

Among other things, ETS proposes to create an online survey to gather feedback from examinees about test site conditions and overall testing environment. Analysis of the survey data will be used to inform our recommendations for enhancing and improving administration materials and test site administration processes. As part of the analysis of the pilot test data, we will summarize the psychometric properties of the anchor pools, including but not limited to providing distributions of classical and item response theory (IRT) item and task parameter estimates, fit statistics, and DIF statistics. One aspect of our evaluation of the pilot test analyses results to inform recommendations for the field test development process will be describing the extent to which the pilot conformed to the designated pilot plan. This survey will provide the Consortium with feedback from the pilot such as the administration of performance tasks.

2.9 Design Item and Student Sampling Plan and Select Anchor Items and Tasks for Calibration and Building the Vertical Scale

The main purpose of the field test is calibrate all of the available items on large-scale samples under near-operational conditions in preparation for the operational program, and to validate, adjust, and strengthen the horizontal and vertical scales established based on the pilot test data. We propose to conceptualize the field test as having two phases. In the first phase, large numbers of students will receive linear fixed forms that are reflective of the overall test blueprint consisting of an anchor set of items that remained unchanged from the pilot test, and a field test set composed of new or revised items, along with one or more performance tasks.
These data will be used to obtain operational item parameters for the pool of available items; evaluate the performance tasks; and validate, adjust, and strengthen the scales. In the second phase of field testing, large numbers of students will receive computer adaptive forms that follow operational blueprints, thereby allowing the Consortium to test the online delivery platform for bugs, help assure that the algorithm and item pools are working appropriately, and verify that the platform is functioning as intended.

2.10 Final Field Test Forms Verification

We will conduct analyses of simulated and operational field test adaptive administrations to monitor the functioning of the adaptive system and of each item pool used for operational field testing. At ETS, we have a long history implementing, administering, and monitoring item-level adaptive testing and will use this experience in working with the Consortium. From 1994 to 2011, we have administered nearly 15 million high-stakes item-level adaptive tests, and we have evaluated thousands of adaptive item pools. This experience has been invaluable in establishing various evaluation methods — the key to this and any evaluation is the set of criteria by which results are judged.

In order to evaluate the adaptive system, we will arrange to obtain all information needed to evaluate the delivery system (e.g., item response strings, resulting ability estimates produced, criteria, and specifications or constraints that the adaptive algorithm is attempting to meet), as well as the composition of each item pool used during field testing from the adaptive delivery Contractor.

Simulations will play an important role in evaluating the operational adaptive algorithm and delivery system. ETS’s simulations will be specific to the Consortium, using item parameter estimates from the field test items and simulated test taker populations that are representative of the students of the Consortium member states. We will work with the Consortium and its technical advisors to finalize the design of the simulation study, but as a starting point, we propose simulating 1,000 examinees at a given number of ability (theta) values (e.g., 15) equally spaced between -3 and +3 and running each simulee through the adaptive algorithm. We can then summarize the results of those 15,000 test events and resulting ability estimates per item pool to examine the degree to which the algorithm and resulting scores meet the criteria outlined below.

While simulations are informative and convenient to conduct, they only provide one source of evaluation data. There is always a risk that the simulations may not adequately predict what happens when real students are administered operational tests. For that reason, wherever possible, we will evaluate results from real students tested during field testing, as well as from simulated data.
2.11 Conduct Psychometric Analysis to Support Field Test Data Review

In Section 2.7, we proposed a detailed set of analyses designed to evaluate the performance of test items, performance tasks, and forms for use following the pilot test. We propose that, in order to take advantage of the investment into the pilot test processes, systems, and report templates, we leverage as much as possible for the field test analyses, building upon and improving on this established foundation. As we note in Section 2.7, there is wide array of software available that might be used to meet the Consortium’s needs. Section 2.7 provides detailed descriptions of all item and performance task analyses that we will perform to support the field test data review process. In Section 2.11, we list and briefly describe the primary set of analyses and the activities that we will complete in support of field test data review meetings and the loading of data to the item bank.

We recommend involving educators from across the Consortium in a thorough review of all pilot and field test items and performance tasks, including information on the performance of items under different conditions (e.g., paper and pencil, accommodated, and translated versions). To achieve this goal, we propose to conduct the field test data review in a similar manner as the pilot test data review (distributed system, online and/or face-to-face, and virtual meetings).

2.12 Conduct Psychometric Analysis to Support Item and Task Calibration

The details of the design, sampling, and the calibration and linking plans are described in detail in Section 2.9. In that section, we proposed to use the method and IRT model that was identified during the pilot test stage; however, if our proposed hybrid design is accepted, even at the field test stage we will have the flexibility to modify the methods if necessary. That is, with the hybrid design we can calibrate and link the forms to the scales using both the common-item anchor (because the fixed linear forms and performance tasks will have overlapping content) and the equivalent groups design. As described previously, this calibration/linking design should be stronger than either common items or equivalent groups alone. However, if any part of the design cannot be met in the data collection phase, we can use one or the other methods.

At ETS, we are experienced in conducting analyses and turning around results in timeframes as short as 48 hours, and we are confident that with the combination of the right processes with the right number and type of staff, we will meet the expected turnaround time. In addition, we have extensive experience in documenting data processing procedures in enough detail for others to reproduce results. On several K–12 contracts, we have worked cooperatively with various independent Contractors and external auditors who parallel process our psychometric analyses to verify the results (e.g., AES, HUMRRO).
2.13 Present to TAC Meetings as Required

We understand that one of the important ways we must support the SBAC Consortium States is through helping them derive consensus on psychometric designs, issues, and decisions among Technical Advisory Committee (TAC) members, the Executive Committee, critical stakeholders such as the U. S. Department of Education, and the SBAC membership at large. Our program manager and the appropriate psychometric and research staff will work with the SBAC to identify contract-related TAC meeting agenda topics in advance of the meeting and to prepare briefing materials to be sent to the TAC members at least 10 days prior to each meeting. In addition, we will work with the Consortium staff to prepare and deliver meeting presentations, carefully listen to TAC discussions, and contribute to the discussions as needed and appropriate. If desired, our staff will take notes and provide these to the Consortium staff.
A. Project Approach/Methodology
A. Project Approach/Methodology

In this section, we provide a complete description of our proposed approach and methodology, and we attempt to convey our understanding of the project. We have described the approach according to each of the 13 main tasks listed in the table on page 17 of the RFP.

2.1 Determine Linking and Equating Design for Special Needs Forms

2.1.1 Pilot Designs

A. Linking and Equating Design

Any data collection design is necessarily a compromise among cost, practicality, and the expected quality of results. In general, one seeks designs that minimize sample size requirements without unduly impacting the quality of results. Designs that are robust to common sources of errors but which remain practical to implement are also preferred. The design options presented here are intended to best balance these considerations.

In theory, the “ideal” data collection design would administer to each student all of the items and tasks to be scaled. However, such a design is neither practical (in terms of what is required of each student) or likely to yield acceptable results (due to the fatigue inherent in responding to hundreds of items and tasks). The solution is to employ a “matrix” design in which each student is administered only a small fraction of the available items and tasks. However, such a design must incorporate some means of linking together onto a common scale the items and tasks administered to different student groups.

Two methods of linking are commonly used. The first is termed “common items” and requires that the blocks of items and tasks be administered to different student samples in part overlap. The second approach is termed “equivalent groups” wherein the test material presented to different student samples is considered as comparably “on scale” by virtue of the equivalence of the groups. Both approaches have advantages and weaknesses. For example, the common-item approach is both relatively inefficient (due to the overlap or redundancy in test material across groups) and dependent on the common items performing equivalently across groups (item position and context effects may cause them not to do so). However, common items are also capable of providing very strong linking bonds. In contrast, the equivalent-groups method is efficient but vulnerable to sampling failures.

Because neither linking method is guaranteed to work perfectly in practice, the designs that we propose incorporate both. This is done by assembling partially overlapping blocks of test content and randomly assigning those blocks to students. The result is a design that is both reasonably efficient and robust to many potential sources of error. The data returned are also ideally structured for item response theory (IRT) calibration.
The data demands for a calibration pilot study are relatively modest. However, if representation from all the Consortium governing states in the pilot is required, sample size requirements can be adjusted accordingly. Some over-sampling of English learners, students requiring accommodations, and other demographic groups will also likely be necessary.

We therefore propose a matrix data collection design that uses both common-item and equivalent groups links to bind all items and tasks at a given grade level to a common scale. The design also incorporates common-item links between grade levels to establish a vertical scale. These links are implemented by administering at most grade levels blocks of test content sampled from the immediately lower grade level. Administering out-of-level content at lower but not higher grade levels minimizes concerns regarding opportunity to learn.

An important element in a matrix design is construction of the forms presented to each student. Because forms will be assembled according to specifications that are determined outside of this proposal, it is useful to list the assumptions made by the data collection design:

1. All test content (summative items and performance tasks) that will be piloted at each grade level will be divided into some number of blocks or forms.
2. Each form will be assembled to represent the test blueprint or content specifications that will eventually drive selection and delivery of the operational adaptive tests. This assumption is required to properly define the proficiency metric against which items and tasks will be calibrated by IRT.
3. Certain pairs of forms will overlap partially with one another by containing some proportion of items in common (depending on the length of the forms and the nature of the items, 10 percent to 20 percent overlap would be targeted).
4. Each form will be of a size that can be reasonably administered in roughly two hours of testing time. Administration could be divided into two sessions, with one presenting a set of summative items while the other administers one or more performance tasks. The sessions could be scheduled across two days, paralleling the operational model where students are presented the performance tasks several weeks in advance of the summative computer adaptive tests. It should be noted that an alternative design is also proposed under which most students are presented either a block of summative items or one or more performance tasks. This alternative is quite a bit more complicated in its particulars than the design proposed, but has the benefit of lessening the testing demands placed on most students and schools.

Because each form is assembled to the same specifications, the same scores and subscores are reportable from each. The distribution of forms to students will take full advantage of the flexibility inherent in online test administration for varying the test content presented to each student. Although it is often inconvenient to spiral numerous paper forms across students in a classroom, doing so is relatively straightforward online.
The basic structure of the design is depicted by the annotated figure below. For completeness, this includes the special (fixed) forms described in Section 2.1.2 of the proposal. Because this figure appears quite detailed at first glance, it is important to distinguish and describe most of elements. **Columns** in the figure represent samples or student groups. Given the spiraling of test content across students, each column or sample is comprised of students from a wide variety of classrooms, schools, districts, and states. **Rows** in the figure represent test content. To better describe the way the forms presented each student are formed, blocks of items are distinguished from sets of performance tasks. The annotations attached to the figure describe many of the key features of the design.
Figure 1. Pilot Data Collection for Linking Design 1  The basic structure of design depicts how forms comprised of item sets and performance tasks can be administered to on-grade and cross-grade student samples as well as students testing on paper and for other special form samples (ELLs and students with special needs).

Notes:

1. On-grade samples of students are presented only test content targeted at their grade level.
2. All students in this design are administered both a form containing blocks of items and one or more performance tasks (PT). Testing may be conducted across several days, paralleling the operational administration model. Forms of summative items are assembled to partially overlap with one another, providing common-item links across forms within grade. Each form is administered alongside one or more performance tasks, with each form-task pair randomly spiraled to equivalent samples of 500 students. A total of about 1,500 students will therefore respond to each form or task, with that sample divided among several distinct pairings. Selected form-task pairs are administered to larger samples to support the required dimensionality analyses. This is indicated in the diagram by the same form being paired with the same task for multiple sample blocks.
3. Students in cross-level samples are presented content from both their grade level (green blocks) and the grade immediately below (blue blocks). The data from these students are used to construct the vertical scale. Forms presented to cross-level samples are assembled of content from both grade levels being linked. This is indicated in the diagram by bars of blue blocks of content from the grade below joined with green blocks designating the content of the current grade level. Although the cross-level forms contain content from two grades, their length roughly

---

**Notes:**

1. On-grade samples of students are presented only test content targeted at their grade level.
2. All students in this design are administered both a form containing blocks of items and one or more performance tasks (PT). Testing may be conducted across several days, paralleling the operational administration model. Forms of summative items are assembled to partially overlap with one another, providing common-item links across forms within grade. Each form is administered alongside one or more performance tasks, with each form-task pair randomly spiraled to equivalent samples of 500 students. A total of about 1,500 students will therefore respond to each form or task, with that sample divided among several distinct pairings. Selected form-task pairs are administered to larger samples to support the required dimensionality analyses. This is indicated in the diagram by the same form being paired with the same task for multiple sample blocks.
3. Students in cross-level samples are presented content from both their grade level (green blocks) and the grade immediately below (blue blocks). The data from these students are used to construct the vertical scale. Forms presented to cross-level samples are assembled of content from both grade levels being linked. This is indicated in the diagram by bars of blue blocks of content from the grade below joined with green blocks designating the content of the current grade level. Although the cross-level forms contain content from two grades, their length roughly
equals that of the on-grade forms. They would ideally be assembled by combining mostly items from the grade below with some from the current grade. Whether the performance task(s) administered to cross-level students could similarly pair tasks from both grade levels (as shown by the red bars with one green block attached) would depend on practical considerations of testing time.

Cross-level forms are spiraled alongside the on-grade forms, to equivalent student samples. This allows the vertical scale to be based on equivalent groups as well as common-item links.

Paper samples will be independent of (and so not randomly equivalent to) the on-grade and cross-level spirals. However, they should be sampled from the same population.

Paper forms would ideally be assembled from intact on-grade forms, allowing common-item comparisons of student performance.

Paper forms will be administered to one classroom per tested grade in sampled schools to yield roughly 1,500 students per form. Forms will be spiraled at the student level within the classroom.

Samples for the special needs forms would also be independent of the on-grade and cross-level spirals. However, given the purpose of the special needs forms, samples will likely be drawn from a distinct student population. Special needs forms would ideally be constructed as versions of the paper forms.

Special needs forms would be administered to samples of roughly 1,500 students each.

The proposed design has a number of both psychometric and practical advantages:

1. The use of both common-item and equivalent-groups links across forms lends the design a unique combination of efficiency, strength, and tolerance to error.

2. The design maximally leverages the capability of online administration to distribute different test content to different students.

3. The design will flexibly adapt to a variety of modifications in both the assembly of test content and the recruitment and sampling of participants. Such flexibility is required in a large-scale pilot, where the realities of the real world rarely permit things to proceed operationally exactly according to plan.

4. The design will provide data that fully support the wide range of analyses required to answer an equally wide range of practical questions.

5. The design will return data that are ideally suited to calibration by a variety of IRT models and calibration software packages.

As described above, the requirement is that all students are administered both a form of items and a performance task component. While this is an advantage in terms of strength of the linking, it can also be a disadvantage in terms of burden to teachers and students. If the testing time per student is impractical, then we offer a second design (please see Figure 2) as an alternative. Under this design, most students are administered either a form or a performance
task component. The general compromise is that the design in Figure 2 requires more samples
and students, but is intended to reduce the testing time required for most students by roughly
half. This design also contains a subsample of students taking both the summative and
performance task components to inform the dimensionality analyses, and has the same
psychometric and practical advantages described above. The same representations are used in
Figure 2 as in Figure 1: Blue boxes designate grade 3 forms, red boxes designate grade 3
performance tasks, and green boxes designate grade 4 content.

**Figure 2. Pilot Data Collection Linking Design 2** The second sampling design is intended to
reduce testing time for most students by roughly half by administering to students either a form
or a performance task; however, this design requires more samples and students.

1. On-grade samples of students are presented only test content targeted at their grade
level.

Most students are administered either a form of summative items or one or more
performance tasks (PT). Testing of each student would likely be completed in a single,
one-hour session. Forms of summative items are assembled to partially overlap with one
another, providing common-item links across forms. Each form and PT is randomly
spiraled to equivalent samples of 1,500 students.

2. To support the dimensionality analyses, selected samples will be administered both a
form of summative items and a PT. Testing within these samples may be conducted
across several days, paralleling the operational administration model. Three to five form-
task pairings would be assembled, with each administered to 1,500 students.

3. All students in the cross-level sample would be administered both a summative form and
one or more tasks. Students in cross-level samples are presented content from both their
grade level (green blocks) and the grade immediately below (blue blocks). The data from these students are used to construct the vertical scale. Forms presented to cross-level samples are assembled of content from both grade levels being linked. This is indicated in the diagram by bars of blue blocks of content from the grade below joined with green blocks designating the content of the current level. Although the cross-level forms contain content from two grades, their length roughly equals that of the on-grade forms. They would ideally be assembled by combining mostly items from the grade below with some from the current grade. Whether the performance task(s) administered to cross-level students could similarly pair tasks from each grade level (as shown by the red bars with one green block attached) would depend on practical considerations of testing time.

Cross-level forms are spiraled alongside the on-grade forms, to equivalent student samples. This allows the vertical scale to be based on equivalent groups as well as common-item links.

Paper samples will be independent of (and so not randomly equivalent to) the on-grade and cross-level spirals. However, they should be sampled from the same population.

Paper forms would ideally be assembled from intact on-grade forms, allowing common-item comparisons of student performance. These forms will be administered to one classroom per tested grade in sampled schools to yield roughly 1,500 students per form. Forms will be spiraled at the student level within the classroom.

Special needs samples would also be independent of the on-grade and cross-level spirals. However, given the purpose of the special needs forms, samples will likely be drawn from a distinct student population. Special needs forms would ideally be constructed as modified versions of the paper forms, and would be administered to samples of roughly 1,500 students each.

B. Sampling of students

Given the stakes associated with eventual SBAC score use, it is critical that item and test statistics are obtained from samples that will enable generalization across the diverse population of students to be assessed, so as to develop fair, psychometrically sound, high-quality assessment instruments that yield accurate and reliable results from which to draw valid inferences.

As noted in the RFP, pilot forms will be administered to samples of students from representative geographic, demographic, and special needs groups. While there may be policy aspects to number and proportion of students drawn from each state and subgroup therein, there are psychometric implications that must be considered. Below we propose, for the Consortium’s consideration, a sampling framework to support the psychometric integrity of the pilot results. We look forward to working with the Consortium and its Technical Advisory Committee (TAC) on further details of the proposed plan or modifications thereof that yield representative samples while balancing practical constraints.
The states that make up the Consortium differ substantially with respect to size and population density, variance in instructional emphasis, and demographic variables associated with educational achievement. Moreover, each state comes to the Consortium with fully developed state-specific curricula and assessment systems currently in use for accountability purposes under the No Child Left Behind (NCLB) Act. For the initial years of the SBAC assessments, the students within the Consortium may actually comprise sufficiently distinct subgroups based on degree of statewide progress in transition to and implementation of the Common Core State Standards (CCSS).

To support scaling and dimensionality studies, the sampling plan must provide representative samples of approximately 1,500 examinees per item/performance task, plus sufficient number of students with disabilities (SWDs) and English learners (ELs) for special analyses and the comparability study.

Given practical limitations on sample size in the context of a pilot, what is the preferred approach to achieve adequate representation, such as different types of student disabilities? What is the preferred weighting scheme for sampling?

We propose that a stratified multistage probability sampling method be employed to select representative samples of students based on two possible policy decisions — equal state representation versus representation proportional to state population (Cochran, 1977).

In this approach, the population, formed by the participating states, is stratified by states or aggregates of states. Within each stratum, a two-stage probability sampling approach will be employed to obtain a student sample. Within states, the determination of additional strata will be governed by policy decisions. Stratification variables of interest may include urban/suburban/rural classification, district/school size, school grade configurations, public/charter, special schools, and demographic distributions (race/ethnicity, SES, EL distribution, SWD). The sampling frames will be formed for each stratum.

In the first stage of selection, schools, the Primary Sampling Unit (PSU), are selected with Probability Proportionate to the Size (PPS) from each stratum. Ideally, there is a second stage of selection, in which a class of students within each sampled school will be randomly selected. Because the number of students sampled from each school is approximately equal, this method will select the students with approximately equal probability.

Slightly different assumptions will be necessary for the pilot of pencil and paper forms with respect to sampling. Online administrations are the primary testing mode; however, we recognize that important differences may exist between schools that have technology infrastructure and those that lack it, and these differences may require some adjustment in the sampling design.
We recognize there may be challenges in obtaining school participation. We therefore propose that sampled schools administer the pilot in every tested grade. In addition, should random selection of classrooms prove unworkable, we propose to provide schools with guidelines for identifying classes to participate.

In analyzing sample data, weights are sometimes applied in estimating IRT models. Several programs such as the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA) have applied weights in the context of IRT calibrations. Typically, this has not been implemented in K–12 programs. However, the viability of this technology might be considered by the Consortium and TAC.

C. Description of how the equating results will be evaluated

Careful evaluation of analysis results is essential prior to, during, and after IRT scaling to examine the integrity of the vertical scales, as well as the quality of the equating of forms within grade. A fundamental aspect of evaluating a vertical scale is to investigate how examinee scores change over grade levels by assessing changes in means and standard deviations of scale scores as well as changes in medians and selected percentile ranks. Prior to conducting the IRT vertical scaling, classical test statistics are examined to evaluate the ordinality of linking items. These statistics might include:

1. Means and standard deviation for all grade levels
2. Medians as well as 10th, 25th, 75th, and 90th percentiles for all grade levels
3. Frequency, relative frequency, and relative cumulative frequency distributions by grade level
4. Number and percentage of examinees obtaining lowest and highest scores at each level
5. Means and standard deviation by school types
6. Examinee demographic characteristics

In general, we would expect to see higher performance for higher grade levels on the cross-level common items. During IRT scaling, model-data fit and convergence criteria for each calibration run will be monitored.

After IRT scaling, we can compare the equated score distributions across grades to see if the vertical scaling results appear reasonable. An important component of evaluating the reasonableness of vertical scaling results is to examine the test characteristics curves (TCCs) across all levels of a content area to assess whether these curves match expectations given the particular content area, the way in which the tests were constructed, and how students typically learn a particular content area. The TCCs for all forms on the new vertical scale, within a content area, will be plotted and examined for overall progression in difficulty. It is expected that, in general, forms will increase in difficulty as grade level increases. This progression in the order of the TCCs on the vertical scale should coincide with the changes in test difficulty over grade level. Likewise, the conditional standard errors (CSEMs) for all tests on a vertical scale will also be
plotted and examined for overall ordinality. It is expected that as test level increases, the ability level at which the test is measuring most accurately will increase as well. Figures 3.a and 3.b show examples of how we would expect TCCs and CSEMs plots to look for a properly functioning vertical scale.

**Figure 3.a: Sample Mathematics Plots of TCCs.** Plots of test characteristics curves (TCCs) allow psychometricians evaluate the ordinality of test forms across grades, an important characteristic for a vertical scale.
Figure 3.b: Sample Mathematics Plots of CSEMs. Plots of conditional standard errors of measurement (CSEMs) allow psychometricians to evaluate that the vertical scale is properly functioning.

To evaluate the horizontal equating, we will also examine various classical test theory statistics, as well as those listed above. For the on-grade common items administered to randomly equivalent groups, we would expect to see very similar performance for the common items. Any variation seen will determine the variation considered within the limits of sampling variability. After scaling, we will examine the equated score distributions to determine whether they differ beyond what would be expected for sampling variation.

**D. Evaluation of Performance Tasks and Associated Scoring Processes**

The pilot test design also allows for the evaluation of the performance tasks and associated scoring processes. We propose to conduct a set of psychometric studies to evaluate the performance tasks, which will be used for the data review meetings described in Section 2.6 to evaluate the items and make recommendations for writing improvements. These statistics will also be uploaded to the Consortium item bank. All of the data collection designs described in this proposal will yield sufficient numbers of examinees to complete the classical item analysis, differential item functioning (DIF), item response theory (IRT) analysis, rater analysis, trend studies, and generalizability analysis, which are described in detail in Section 2.7.
In concert with the Consortium and their scoring Contractor, we will also provide technical support for the scoring process as deemed necessary, including range finding activities and calibration paper selection. We have experience working in various capacities to support constructed-response scoring, both as the scoring service provider for large-scale assessment programs and in collaboration with other contractors that provide scoring services. To verify the accurate rendering of images that reflect the properties of a student’s response to be scored, we propose to conduct small-scale studies, in collaboration with the scoring Contractor, to evaluate image quality, such as gray scale, can affect what is viewed by raters and can lead to difficulty in score assignment. These studies can be conducted using test forms with known properties (i.e., light writing) to test properties of the system. This approach is much like a “test deck” used to verify multiple-choice scoring. We will work with the scoring Contractor to include a description of how optical images of student responses are distributed to raters and to include the results of these analyses in the technical documentation. In terms of the distributed scoring system, we look forward to discussing these studies with members of the Technical Advisory Committee (TAC) to verify that scores are consistent across jurisdictions.

### 2.1.2 Special (Fixed) Forms

We recognize that the Consortium tests will need to include paper and pencil forms for students without access to electronic platforms and special needs forms. These fixed forms will satisfy the content coverage, cognitive complexity, and item type requirements and will be linked to the vertical scale so that they can be interpreted using the same scale as the adaptive tests. In order to place the paper and pencil and the special needs forms onto the vertical scale established for the adaptive test, a special data collection and study will be completed during the pilot test. We propose to use online forms for the paper and pencil and special needs forms, with the goal of having the majority of the items administered in both online and paper and pencil modes. This will allow for a common-item linking design to place the fixed forms on the vertical scale and will provide the data for the comparability study described in Section 2.7. The text below describes our proposed approach.

**Form Selection**

To begin, we will work with the Consortium to identify online forms that could serve as paper and pencil versions. Our goal will be to identify forms that will perform as similarly as possible under all conditions. We recognize that there may be some innovative or technology enhanced items that will not be convertible to paper and pencil. In these cases, we will consider replacing those items with items that are convertible that measure the same content and cognitive complexity. We will want to involve the Consortium and their advisory committees in the selection and approval of these forms. Those same forms will then be used to create the special needs forms. We are proposing to use the same paper and pencil forms for those students that require translations, accommodations, or modifications because we suspect that the samples of these students may be quite small in the pilot study, and this approach affords us a way to link the forms to the scale even if the sample sizes are too small to complete a separate study.
**Sampling**

Once identified, we recommend administering these intact forms along with designated performance tasks to samples of approximately 1,500 students each, which are as similar to the online sample as possible. One option to obtain this type of sample would be if a subset of schools could identify two classrooms per grade to be tested — one for online administration and one for paper administration. It would be ideal if students from these classrooms could then be randomly assigned to a testing mode. However, recognizing that this might not be practical, another approach would be to identify matching schools and assign one school from each matched pair to either the online or paper and pencil version. For the special versions of the paper and pencil forms, samples will be drawn from the distinct student populations with the goal of obtaining as many cases as possible, up to 1,500 cases per form. We will make these special needs forms available to all schools participating in the online and paper pilot administrations. Additional sampling of schools with historically high use of the special needs forms of interest to the Consortium may be required to reach target case counts for comparability analysis. However, we also recognize that it may be impossible to obtain a sample of this size for these test forms. In that case, we will link these forms through the paper and pencil forms, in one of three ways, which we describe below.

**Linking Design**

There are several approaches for putting these forms on the vertical scale. The choice of which approach depends in part on policy considerations, and in part on the nature of the samples obtained for the paper and pencil and the special needs versions of the forms.

The most straightforward approach is to include all students in the calibration and vertical scaling studies (please see Section 2.4), including those who took the paper and pencil or specials needs form. The resulting item parameter estimates will already be on the vertical scale. The resulting scores will be on the same scale as the adaptive tests.

If students taking the paper and pencil and the special needs versions of the forms are excluded from the calibration and vertical scaling studies, then two approaches exist.

One approach is to simply use the parameter estimates that result from the calibration and vertical scaling study completed with only the online samples. This approach declares the supremacy of the online administration mode and assumes invariance across administration modes and samples, despite any differences in converting items to paper and pencil and/or the special needs.

The second approach, and the one we recommend under this condition, is to separately calibrate the paper and pencil forms with the associated sample, and then horizontally link these forms to their online equivalent using the entire test form as the anchor. Common items that performed differently by mode could be excluded from the anchor set. This approach
adjusts for differences in the item parameter estimates resulting from any difference in the administration method (online versus paper and pencil). This same approach could be used for the special needs forms, sample size permitting. However, despite over-sampling, it is quite likely that there would not be sufficient numbers of students to perform separate calibrations for this group. In this situation, because the special needs versions are the same as the paper and pencil forms, we recommend including these cases in the paper and pencil sample for the calibration and linking studies.

Comparability between Fixed and Adaptive Forms

Operationally special fixed form tests will be linked to the operational vertical scale so that reported scores are on the scale as the adaptive tests. When items are administered in different modalities (online versus paper and pencil) or when they are translated or otherwise transformed, it is not unusual for some items to perform differently. Administration of fixed forms during the pilot provides an opportunity to evaluate the impact of paper and pencil representation and other transformations at the item and test level. To this end, we will conduct comparability analyses of the fixed forms administered during the pilot stage, and prepare a report for the Consortium for consideration. At the item level, these analyses will include differential item functioning (DIF) analyses, scatterplots by mode of standardized classical item difficulties, and a comparison of item parameter estimates and item fit statistics. At the test level, we will make initial comparisons of score distributions in paper and online modes. We will also examine the summary statistics (mean, standard deviation, median), distribution of scores, the test characteristic curves, and the conditional standard error of measurement, as well as the reliability estimates of the forms under each of the conditions. We will also administer a computer familiarity questionnaire to both groups (Kirsh, Jamieson, Taylor, & Eignor, 1998); this will include questions about computer access, use, and experience, as well as test preparation. The information from this questionnaire will be summarized and will help interpret the results of the data collected. For a detailed description of the analyses, please see Section 2.7.

2.1.3 The contractor will provide a detailed written description of all planned psychometric activities with sufficient time for SBAC staff and committees as appropriate to review and approve the plans.

One of the keys to a successful project is clear communication and detailed planning of all planned psychometric activities. Due to the ground-breaking work that is being undertaken by the Consortium, it is of paramount importance that detailed descriptions of the designs, implementation and analysis plans, and evaluation criteria be documented as well as reviewed and approved by the SBAC staff and committees. We expect that we will be working with not only the Consortium, but also with various other vendors, so clarifying the nature of the plans — along with the key deliverables, requirements from the Consortium and their other vendors, and all handoffs among the vendors — is important. While we will work with the Consortium on establishing the time associated with the review, we have planned for a minimum of a 10-day
review period by the Consortium staff and committees for major documents and plans. However, we also know that given the magnitude and significance of this project, we will need to work closely with the Consortium staff and committees in determining the right amount of review time. We’re flexible in this regard and are prepared to provide more time, or at the direction of the Consortium, work in a “just-in-time” manner.

2.1.4 Twice Annual Contractor Meetings

A key component of effective collaboration between ETS and the Consortium will be our twice-annual contractor meetings. These twice-annual contractor meetings will cover a broad scope of activities including potential cross-Consortium contractor activities. ETS will work with the Consortium in the planning of these meetings to establish that the contractor meetings satisfy the needs of the Consortium, other collaborators, and the ETS/Measured Progress Collaborative. The ETS Program Manager will collaborate with the Consortium to facilitate these sessions. We anticipate that one meeting will occur during the first several weeks of each program year, and the second meeting will occur at an appropriate mid-year point.

For each of the contractor meetings, ETS will provide a draft agenda at least 10 days prior to each meeting, to allow the Consortium to review, comment, and add additional agenda topics. Preparation for meeting agenda topics will be assigned to specific contractor personnel involved in providing the psychometric services and completing the deliverables. Program Management will assume the responsibility to record minutes of the meeting. The minutes will address all tasks (with particular emphasis on questions or issues that need to be resolved), contract implementation timelines, and decisions reached. Meeting minutes will be published and available by end-of-day three business days after meeting adjournment.

We propose using the first annual contractor meeting within a few weeks of contract award and the estimated start date of January 9, 2012, as the Launch/Planning meeting for the entire effort to provide psychometric services and deliverables for the Consortium summative and interim assessments. Depending on the Consortium attendees at this meeting, we propose this meeting be held at a central location for the Consortium, such as Chicago or Denver. ETS in-person staff at this planning meeting will include the Program Executive Director, Program Manager, Lead Psychometrician, and one additional Psychometrician. Other attendees with specific subject-matter expertise or that the Consortium requests will be available to teleconference in as needed. The agenda will cover each RFP task so that details about the Consortium’s expectations and our proposed methods of execution are agreed upon and clarified. We will come prepared with a draft Microsoft Project Schedule that can be used as a baseline for discussions of timelines. A result of this meeting will be an Initial Work Plan for the Consortium to review and comment on, leading us toward a final posting of the Work Plan and a final, agreed upon baseline schedule for the project.

We will work with the Consortium to plan and schedule the remaining meetings to meet the needs of the project, but suggest that the subsequent contractor meetings be timed in
according to when the evaluation/revision of the work plans are due to allow for group discussion and collaboration in the creation of the revised plans. We again propose these meetings take place in a mutually agreed upon central location, such as Denver or Chicago. Suggested timeframes for these meetings are:

- June 2012
- November 2012
- May 2013
- November 2013
- May 2014

ETS will collaborate with the Consortium in a similar manner to plan, conduct, and report each of these subsequent contractor meetings. These meetings will allow the ETS/Measured Progress Collaborative team to review accomplishments as well as candidly discuss future work plans and challenges, issues/concerns, schedules, commitments, Consortium expectations, and any required modifications to the plans and schedules. We understand the importance of relating the work and discussions to the Consortium’s broader goals and efforts in other areas that are addressed by separate contracts. We also recognize the need for the Consortium’s criteria for proficiency to be developed in concert with and comparable to criteria used by PARCC.

### 2.2 Develop Technical Manuals

We will deliver an on-time and accurate SBAC Technical Manual. We are dedicated to working with the Consortium and its Technical Advisory Committee (TAC) to design comprehensive documentation that will provide evidence to support the validity of all aspects of the SBAC pilot and field tests. This document will serve as the ultimate source of information about the SBAC assessment program and will be designed to serve the needs of various educators and stakeholders of the Consortium. For that reason, we are envisioning one master document that is essentially a compilation of various stand-alone technical reports. That way, constituents can refer to and reference a certain chapter of the Technical Manual without having to read the entire document or refer to other sections for context. Of course, the primary goal of the Technical Manual is to meet the needs of the Consortium, and as such, we will collaborate to design the Technical Manual(s) to the exact specifications desired by the Consortium and its technical advisors.

The Technical Manuals will include all aspects of the Consortium summative tests including performance tasks, as well as information about the interim assessment components, from development and technical designs through implementation and item-, task-, and test-specific information corresponding to the pilot and field tests. We will include all of the elements listed in Section 2.2 of the RFP in the Technical Manuals.
The Technical Manuals will:

» include the information needed to meet all current and future USED peer review guidelines

» contain detailed descriptions of all procedures or analyses used to evaluate pilot and field test results

» be organized and labeled to permit easy cross-referencing to the Standards of Educational and Psychological Testing (AERA, APA, & NCME, 1999)

In addition to being comprehensive, the Technical Manuals should also be user-friendly. For that reason, we propose creating an online version of the Technical Manuals in addition to any hard-copy or electronic (PDF) versions as required. The Consortium can then post the Technical Manuals on its website. Given the length of the document(s), a web version will make widespread distribution of the report simpler and more cost-effective. In addition, hot-linked tables and other terms within the text will facilitate searching, making the navigation of the report and finding relevant information quicker and easier. The Consortium may also be interested in developing a technical documentation website, with hyperlinks to the Technical Manuals and related documentation, as well as corresponding data access. Should that be of interest, we will draw on our experience with the NAEP technical documentation website.

The Technical Manuals will undergo an internal review process to verify accuracy and completeness. This review process will also verify that style, format, and content are consistent with the Consortium’s standards. The Technical Manual will be reviewed and modified based on input from the Consortium and its technical advisors. We will work with the Consortium to help establish that the document meets the needs of the Consortium, and is of the highest possible quality.

We understand the importance of all documentation being sufficiently complete for a qualified independent contractor to replicate all technical activities. In addition to clear and complete text describing each aspect of the program, we propose an appendix to the Technical Manuals that would include specifications for all analyses, including descriptions of the software used for analyses, along with all necessary code or command files needed to replicate the analyses.

We also understand the considerable planning and effort that will be required to assemble information from the various Contractors to help ensure that all aspects of the SBAC system are completely transparent. We will draw upon our direct experience in working successfully with a number of Contractors to create technical manuals for our current clients.

**Documentation of the item development process.** We have extensive experience working with item development Contractors to fully document the item development process. To create that documentation, we will work with the Consortium’s item development Contractor and representatives from the Consortium to outline the steps taken in the creation of the item specifications and item writing materials; the item development and review process; and the
external content, bias and sensitivity, and data reviews. Results of those meetings will be described, including summaries of outcomes from the meetings, such as the number of items reviewed, and the numbers and percents accepted, rejected, or marked for revision. The documentation will also include how the item developers align the items to the Common Core State Standards (CCSS) and how items are modified for accessibility.

**Full documentation of dimensionality analyses.** ETS will include a description of the methodology for the dimensionality analyses in enough detail within this section of the Technical Manual for an independent Contractor to replicate the analyses, including the software used for analyses, as well as the command files to run the analyses. Our recommended approach to assessing dimensionality, including multidimensional item response theory (MIRT) and confirmatory factor analysis, are presented in Section 2.4 for the Consortium’s consideration. We will present results of the selected dimensionality analyses in tables and text, describing the various models and overall fit indices and the implications for the development of the vertical scale.

**Full documentation of model selection analyses.** ETS will explain the model selection process in detail, including description of the methodology and software employed to test various models. As appropriate, we will include cross-validation as an approach. We will present the results of the model-data fit analyses conducted, including chi-square and residual analysis, as well as the pros and cons of the various models under consideration and implications for growth. When evaluating and comparing IRT models, it is important to go beyond statistical measures of fit (which nearly always determine that more complicated models fit better) to examine the practical implications of model misfit. Misfit can have small or large practical implications. For example, misfit that understates the standard error of an ability measurement can have minimal effect on typical score reporting but can have a more serious effect on adaptive testing. We will describe the models selected by the Consortium, along with the rationale for selecting those models and the assumptions associated with those models.

**Item calibration and analysis documentation.** The item calibration and analysis documentation will include step-by-step explanation of the item calibration process, including which models and software were used. This section of the Technical Manual will also include a description of the samples used for calibration and other analyses, including how translated and other accommodated items were calibrated. We will include summary information on the item statistics that are calculated, both classical and IRT, as well as the list of all acceptance criteria for items. We will provide information on the number of items with unacceptable criteria, including the reason(s) that an item was rejected and what content the item covered. We will outline the item review process and quality control procedures for review in detail.

**Documentation of performance task development.** ETS will work with the Contractor chosen to develop performance tasks and to fully describe all steps in the task development process. We will document the lessons learned from item tryouts and cognitive labs and the resulting
revisions to tasks. We will explain all aspects of the field testing of performance tasks, including the design for field testing, process of range-finding and anchor paper selection for rater training, strategy for establishing inter-rater reliability in real-time during field testing, and the analysis of inter-rater reliability results of the field test performance assessments. We will describe rater reliability and trend studies and other types of studies conducted, designed in consultation with the Consortium, along with corresponding statistical analysis. Finally, we will report the method that was used to help establish task difficulty comparability and the criteria each task must meet to be accepted, as well as acceptance rates.

Scaling documentation. The Technical Manual will explain all aspects of the scaling and equating of the assessments, including the measurement model approach, procedures for creating and maintaining the vertical scales, the operational methods of linking, and a detailed description of the sampling design. ETS will organize the sections of this chapter on equating and linking, as well as standard setting, in such a way that each could serve as stand-alone technical reports.

Adaptive test algorithm documentation. We will work with the adaptive testing delivery Contractor to fully explain how the adaptive testing algorithm functions. We have two goals for this section: The first is to document the adaptive testing in enough detail that an independent consultant could replicate the algorithm; and the second is to explain the procedure in a way that makes the details accessible to audiences without expertise in adaptive testing. We want to make the explanation of this aspect of the Consortium entirely transparent and clearly understandable, so that students, teachers, parents, and other stakeholders understand how scores were calculated. This section of the Technical Manual will outline all decisions and constraints imposed on the adaptive system, as listed in the RFP, including:

» entry method (e.g., selection of first item, initial scoring, and method for preventing premature convergence)

» selection criteria (e.g., item selection method, method used to satisfy content constraints, width of informative search [may vary throughout test and affects pool design], and exposure control to maintain item security on the summative assessment, and to help assure that students do not see the same items from one administration of the interim assessment to another)

» scoring (e.g., momentary score estimation, final score estimation, and method of calculating SEM and item information throughout each test and at the end of testing)

» termination (e.g., trigger for test end, and final calculations)

» output (e.g., information reported on screen immediately, and information retained for database and reports)

» Please note that documentation of the studies conducted to evaluate the adaptive testing algorithm and field-testing item pools will be included in the Technical Manual section describing technical information about each summative test.
Accommodations, including differential item functioning (DIF) analyses and comparability analyses. We agree that it is important to disaggregate students by disability subtype when reporting item statistics and performance results, because students with different disabilities may interact with test items and tasks in different ways. The Technical Manual will provide information on the types of accessibility features used, as well as the number of students using specific accommodations, overall and by disability subtype, and English learner (EL) status.

Differential item functioning (DIF). We will describe DIF studies for students with disabilities (SWD) in detail, including definitions of reference and focal groups, and sample sizes used for analysis. As permitted by sample sizes, we will explain the DIF statistics, as well as the various levels of DIF categorization and what they mean. If alternative measures are used to assess differences, then we will present those in this chapter as well. (For example, if equivalence of the test construct across accommodated/nonaccommodated groups and groups that used/did not use accessibility features will be tested with factor analysis and residual analysis.)

Many recent publications have examined the psychometric properties of K–12 assessments including the relationship between testing accommodations and performance, dimensionality, and growth (e.g., Bolt & Ysseldyke, 2006; Buzick, 2011; Buzick & Laitusis, 2010; Cohen, Gregg, & Deng, 2005; Cook, Eignor, Sawaki, Steinberg, & Cline, 2010; Cook, Eignor, Steinberg, Sawaki, & Cline, 2009; Engelhard, 2009; Finch, Barton, & Meyer, 2009; Johnstone, Thompson, Moen, Bolt, & Kato, 2005; Kato, Moen, & Thurlow, 2009; Klockars & Lee, 2008; Laitusis, 2010; Steinberg, Cline, Ling, Cook, & Tognatta, 2009; Stone, Cook, Laitusis, & Cline, 2010).

Comparability. We will report average item scores for students using specific accommodations and accessibility features, and we will compare this to average item scores for students not using accommodations or accessibility features.

Security procedures and integrity of assigning scores to students/schools. ETS will report on documentation concerning security procedures from all Contractors, as applicable. We will likely obtain the procedures for preventing and detecting inappropriate activities (cheating) from the Test Delivery Contractor and/or the Consortium, as well as the steps taken when cheating is suspected. Description of requirements for testing integrity would likely include information on data encryption, prohibiting of certain functionality (e.g., printing, saving, copying, etc.), automatic shutdown during idle periods, clearing of the cache following testing, firewalls, and restriction of staff access via passwords. We would also include descriptions of strategies used to detect cheating, such as any software used to detect similarity across essays or with text available on the Internet. In addition, we will include documentation in this chapter of all security measures taken to protect the summative items and tasks, from development through administration and scoring. This will require gathering documentation from many sources,
including the Item and Test Development Contractors, the Delivery Contractor, the Scoring Contractor, and the Consortium.

We will describe the processes in place at all stages of the test cycle in the Test Security chapter of the Technical Manual, including the following: Test Development, Item and Data Review, Item Banking, Transfer of Forms and Items, Online Security, Printing and Publishing, Test Administration, Test Delivery, Processing and Scoring, Data Management, Transfer of Scores via Secure Data Exchange, Statistical Analysis, Reporting and Posting Results, and Student Confidentiality and Test Results.

**Test development process.** In the Technical Manual, we will document the test development process in partnership with the test development Contractor and representatives from the Consortium. We are experienced at providing similar information in technical reports for a variety of programs and will provide a thorough summary of the process. We will work closely with the test development Contractor to understand and describe every step of the forms construction process, including planning and setting the test purpose; detailed description of the steps or algorithm used to construct forms; description and counts of item types by content and grade; test specifications and design, including target test characteristic, test information and conditional standard error of measurement (CSEM) curves, and test blueprint designs; and all internal Contractor reviews and external committee reviews.

In addition, we will outline the technical criteria for test acceptance in this chapter, including the following: minimum values of test information and CSEM for each test constructed, whether for the adaptive test events or in the fixed forms; validity evidence based on content; correlational evidence; evidence based on internal structure; and reliability evidence overall and by subgroups, including stratified alphas, decision accuracy and consistency measures, and inter-rater reliability for performance tasks. In separate sections of this chapter of the Technical Manual, we will describe the test development process for the pilot and field tests.

**Technical information about each summative test.** ETS will begin this chapter of the Technical Manual by outlining the specifications of the summative test, including the statement of test purpose and the test blueprints by content and grade, as well as descriptive information about testing time, disaggregated by content, grade, and various subgroups of interest. Next, we will describe the pool characteristics and results of the adaptive testing algorithm in detail. This section will include information about the number of times a student can be validly tested from each pool, and pool characteristics and pool adequacy report. We will present full description and results of the simulation studies that show how well examinee ability is recovered, by analyzing estimation bias and root mean square error (RMSE) results. We will use the results from both simulation and field testing data to report reliability and fidelity to test specifications, including an analysis of item pool content composition, the number of test events that meet all specifications, and test information functions of pools, both overall and for subscores. For the
results of the final field testing (using the operational design), we will report reliability and test information functions overall and by subgroups of interest, including accessibility versions.

**Additional information for performance tasks.** ETS will work with the Contractor chosen to develop performance tasks to fully describe all of the methods related to scoring. In addition, we will provide explanation of how performance task scoring elements were verified and summary statistics of student performance on tasks. The development Contractor will need to provide evidence of score consistency across prompts. This will include detailed documentation of scoring procedures, such as description of the scoring design (field-test sampling, real-time score monitoring, and range finding activities) and a report on how these procedures were carried out, with rater consistency results presented by comparing means and standard deviations of rater scores, effect sizes of differences between rater scores, Kappa and weighted-Kappa statistics, and product-moment correlation coefficients. In addition, we will summarize performance on tasks with both descriptive statistics and classical statistics, including results of differential item functioning (DIF) analyses, and item response theory (IRT) statistics. As applicable, information about the validation and performance of any artificial intelligence (AI) scoring used during the pilot or field phases will be included.

**Information about the interim testing system.** ETS will work with the Consortium’s interim system Contractor to obtain the appropriate documentation to describe how the interim system is intended to be used, including explanation of valid uses of the interim system. We will document technical guidelines for using the system here, as well as an explanation of professional development materials provided to system users to encourage valid system use and the resulting interpretation. We will also describe the plan for monitoring item parameter drift and maintaining the stability of the interim system. We recognize the importance for all stakeholders who intend to use the interim system to fully understand its features. We will write the report with the intention of enabling users with various levels of assessment expertise to take advantage of what the system has to offer.

### 2.3 Determine Standard-setting Design

ETS’s collaborator and subcontractor, Measured Progress, will have the lead role in determining the standard-setting design. Measured Progress has extensive experience in conducting standard-setting meetings. Results of our work have been presented at various national professional meetings, and were included in a chapter entitled “Setting Performance Standards Using the Body of Work (BoW) Method” in *Setting Performance Standards: Concepts, Methods, and Perspectives*, edited by Gregory J. Cizek and released in 2001.

We have implemented standard-setting sessions in many states including Kentucky, Maine, Massachusetts, Montana, New Hampshire, Rhode Island, Vermont, Florida, Missouri, Nevada, New York, New Jersey, Utah, South Carolina, and Colorado. Additionally, Measured Progress is currently working with the National Assessment Governing Board (NAGB) to assist with setting
standards for college and career readiness in Reading and Mathematics for the National Assessment of Educational Progress (NAEP). Our work on this contract has allowed us to develop a paperless form of standard setting where all presented materials and data collection are done using an online process. Our involvement in this project will position us to help the SBAC establish its definitions and performance standards for college and career readiness.


ETS has experience working with cross-state standard setting through our development and implementation of a multistate standard-setting process for educator licensure tests (Tannenbaum, 2011), and has previous experience conducting standard setting for computer adaptive tests (O’Neill, Tannenbaum, & Tiffen, 2005). Together, ETS and Measured Progress provide a vast array of experience and expertise in standard setting.

**Our process.** Our process will be designed to not only draw on the collective experience of the ETS/Measured Progress Collaborative, but will include an array of expertise from the larger measurement community. Together, ETS, Measured Progress, the SBAC, and the external consultants will form the standard-setting design committee (SSDC) for the SBAC. As partners, ETS and Measured Progress bring a vast amount of statewide standard-setting experience for general education, alternate and English proficiency assessments at both the national and international level, as well as research in standard-setting design and implementation. Most importantly, our companies have experience designing and implementing standard-setting studies in nontraditional settings, which require research-based, innovative solutions. Our strengths lie in adjusting existing methodologies so that they are tailored to the situation at hand. By working together with members from the larger measurement community, this experience will be further expanded. The SSDC will meet with SBAC officials on three separate occasions, during the beginning, middle, and end of the design process.

The SSDC will provide three added benefits to our design process. First, by including other members from the research community, we will have a broader perspective. This will provide the SBAC with a solution that is driven by a sound research-based foundation and not purely by operational expertise. Secondly, by having the SSDC meet regularly, they will be involved in the decision-making process so that operational design remains true to research-based principles and meets the SBAC’s needs and expectations. Finally, the SSDC will serve to validate our design
and recommend the evidence that should be collected to form and support a comprehensive validity argument for the standard-setting process.

Members of the SSDC will be recognized leaders in the field of psychometrics, and they will have significant expertise in the area of standard setting. SSDC members will be determined upon award of the contract.

**PARCC Comparability**

The Common Core State Standards (CCSS) will serve as the cornerstones to both SBAC and PARCC programs. Students throughout the nation will be assessed on these standards, and as a result there will be a natural tendency to compare performance across the two assessments. In theory, neither the home state of the student or the assessment they take should influence the interpretation of that student’s score. It is for this reason that comparability between the SBAC and PARCC needs to be a driving force in the standard-setting design.

Upon SBAC approval, key PARCC stakeholders will be invited to our SSDC meetings. Participating in our SSDC meetings will further expand the perspective that is brought into the process while broadening our operational considerations. Most importantly, by including PARCC stakeholders, we are in keeping with the collaborative spirit of the education reform initiative, and we are providing the foundation for comparability across the two consortia.

Given that both consortia are building their assessment to measure the CCSS and to have comparable standards, it is necessary for the achievement level descriptors (ALDs) to be consistent. To address this need, we propose following a process for refining the SBAC ALDs.

Once the comparability of the ALDs between the SBAC and PARCC has been established, the comparability of the technical, standard-setting procedures, while still significant, becomes less of a challenge. Towards that end, steps should be taken so that both consortia are using parallel standard-setting methods, that is, methods from the same framework (e.g., item versus person-centered method). Discussions between the two consortia should be ongoing throughout the process. As an additional step, various empirical considerations (e.g., performance trends, percentage of students in each achievement level, success in postsecondary education) should be evaluated for consistency.

**SBAC Technical Advisory Committee**

Measured Progress excels in working in partnerships with clients and their technical advisory committees (TACs). Our goal is to work with TAC members in developing a solution that optimizes the balance between client needs and psychometric rigor. The standard-setting design is scheduled for a two-year period, which is ideal in the sense that we can incorporate TAC feedback and recommendations as part of the process rather than conduct a post-hoc critique. As part of our design process, we will present to the SBAC TAC activity around the standard-
setting design for review and feedback. The SBAC TAC will be particularly helpful in making sure the standard-setting design is consistent with the details of the assessment program and the goals of the SBAC. This feedback and any recommendations will help assure technical adequacy and PARCC comparability, and this will be our goal in working collaboratively with the SBAC TAC.

**Panelist Selection**

As was emphasized in Cizek and Bunch (2007), the particular group of panelists selected is an aspect of standard setting that can affect the recommended standards. Given the criticality of the panelist selection to the validity of standard-setting results, the manner by which panelists are selected and the representation of said panelists is of paramount importance. Working collaboratively with the SBAC and the TAC, Measured Progress will develop a comprehensive multistaged approach to select panelists. We will develop methods that will help assure a broad level of representation from teachers. This will include teachers from all grade spans as well as teachers from each member state. We will also carefully consider the representation of non-teacher educators and the general public. The philosophy behind this approach is to find a collection of panelists that will result in the contribution of the collective wisdom of a broadly representative body. We will select panelists such that appropriate demographics (gender, race/ethnicity, and geographic location) are represented, and so that all SBAC states could participate in the process.

Our measurement specialists will work with the SSDC and the TAC to develop and refine our selection process to help assure our goal of broad representation. This approach will also allow for teachers to be integrally involved in the process. Our psychometricians have extensive experience in working with technical advisory committees on a variety of matters, and we understand the role that these committees will have in such a process. We are prepared to present the details of our selection plans and incorporate any feedback that the TAC, the SBAC, or PARCC might have in our process. We will seek approval from the SBAC prior to formalizing any part of the selection and nomination process.

**Number of Panelists**

A goal of any standard-setting study should be to have a reasonable number of panelists such that the results are meaningful and reliable. The guidance provided by the *Standards* (AERA, APA, & NCME, 1999) is that “a sufficiently large and representative group of judges should be involved to provide reasonable assurance that results would not vary greatly if the process were repeated” (p. 54). A number of design elements, including the adaptive nature of the assessment, the desired comparability with PARCC, and the relationship between performance on the assessments and success in higher education, will need to be carefully considered when finalizing the number of panelists. To this end, we may want to consider a distributed standard-setting type of model, where panelists participate remotely in the process. This will allow us to maximize the number of panelists while minimizing costs. By having a larger number of panelists than what might typically occur during a standard-setting study, we will be able to have more
teachers integrally involved in the process. Under this type of model, all SBAC states could potentially participate in the standard-setting process.

**Method**

Below is an outline of main considerations that will be factored into the standard-setting design. As other psychometric details are developed through the remaining sections of this RFP, other considerations will likely be incorporated into the design. This will help assure that our standard-setting design is a foundational element of an integrated assessment system. One of the driving factors in our approach will be the use of customized tools designed to enhance the standard-setting process. Tools that will be outlined as part of our overall design can be developed to perform a variety of functions to support the standard-setting process. For example, panelist recruitment and selection, automation of materials distribution, and collection of panelist feedback and interactive presentation of empirical data.

**Design considerations.** Because of the use of adaptive tests, we will need to sample from the pool of tasks in order to minimize any bias in estimated cut scores, and to optimize the number of tasks (or work samples) that are reviewed by the panelists. Performance tasks will also be included as a critical component in the standard-setting design as the characteristics of the tasks will likely have a significant impact on what is required of the panelists. Another consideration is the use of replicate panels to maximize the number of tasks reviewed and minimize the standard error around the cut scores, and to have built-in procedures for validating the results. The context of distributed standard setting also introduces many design considerations. For example, training panelists; establishing a common understanding of the ALDs; evaluating the knowledge, skills, and ability represented in the tasks; and so forth. By working closely with the SBAC, SSDC, and the TAC, we will optimize the solutions to the various design considerations that we believe will lead to an innovative standard-setting approach that will exceed the needs of the SBAC program.

**Use of empirical data.** In addition to the traditional use of impact data (feedback based on the available SBAC test results), success in college, career training for high school, and in the next grade level all need to be incorporated in the standard-setting design. Given that the prediction of success in college and career training is a fundamental aspect underlying the desired outcomes of the standard-setting process, establishing an understanding of the relationship between performance on the SBAC assessments and these variables will have a critical impact on the standard-setting process. Given the unknown nature of how these variables relate to the SBAC program, it will be necessary to research these relationships and perhaps develop several models that can be used in standard setting. For example, the TAC and SSDC may recommend that we use these data elements early on in the process to establish expectations around the cut scores, or they may request that we use these data as an additional source of feedback to the panelists during the process.
In determining the empirical data used and the particulars around how the data are incorporated into the standard-setting process, some consideration should be given to the data that are being used by PARCC. Different approaches and models, such as these, will in part be driven by the various relationships that are found through carefully conducted research studies and the underlying SBAC policies. Throughout our design process, we intend to outline a series of research studies, develop specific research questions, and design models where the data could be used in the standard setting.

**Panelist agreement.** At a minimum, the basis for panelist agreement should be established through an iterative methodology that allows for multiple panelist discussions between rounds. In the event that we use distributed standard setting, it is likely that we will have a fairly large number of panelists. Given the likely variability in panelist ratings, we may consider alternative statistical techniques, such as the use of outlier analyses, to understand if individual panelists or groups of panelists are having undue influence on the process. The use of statistical techniques in conjunction with comprehensive panelist evaluations may provide the necessary feedback to determine how to best establish panelist agreement and aggregate panelist results.

**Approval Process**

As part of our design process, we recognize that the policy decision-making phase (approval from decision-making bodies) is an integral part of the standard-setting process. Based on this, the design process and the results of the standard-setting judgments will be presented to the executive committee and governing state chiefs for modification and/or approval. Presentation of the results includes, but is not limited to, cut scores, standard errors around cut scores, variation of cut scores and standard errors across rounds, impact data, and panelist evaluations. The results may also require evaluation by the SBAC TAC, SSDC, and at the discretion of the SBAC, any other interested parties.

Our design will specifically address how this approval process will work, who should participate and/or who should be represented, and procedures and protocols, and it will provide guidelines on how to evaluate the results. We envision an approval process that is likely to have a sequence of approvals to help assure that the particular data presented will be tailored to each audience. Our approval process may also suggest the formation of an executive committee that includes key stakeholders from both consortia to further help assure comparability across the assessment programs.

**ALD Refinement**

During the standard-setting meeting, clarification around the language in the ALDs is an important part of the process. This gives panelists the opportunity to operationally define the knowledge, skills, and abilities that correspond to the ALDs. Of particular importance are the knowledge, skills, and abilities that differentiate one level from another. Our standard-setting design will outline a process for capturing these distinctions. Most importantly, our design will show how the panelist recommendations for clarifying the ALDs will be collected, documented,
and submitted for approval by the decision-making bodies. Given the likelihood that a large number of panelists will participate in the process, it may be necessary to develop a collection of web-based tools to collect and summarize the panelists’ ratings. Our design will outline key features of these tools.

Once the standard-setting meetings have concluded and after any adjustments have been made based on policy, it may be necessary to engage in a validation process around the ALDs. Based on the cut scores, it may become necessary to review the alignment between the ALDs and the knowledge, skills, and abilities measured by the corresponding items on the assessment. This could result in a refinement of the ALDs. The standard-setting design will outline various options for how and when the ALD refinement process will take place.

**Benchmarking Against External Assessments**

The cornerstone to gathering evidence to support validity claims will ultimately be grounded in the standard-setting procedures. A second piece of critical evidence can be gathered by benchmarking the resulting standards against other external assessments. This process will allow us to explore the relationship between the SBAC and other assessments such as NAEP, PISA, TIMMS, and of course PARCC. Our approach will show how other data elements (e.g., success in college, career training for high school, and success in the next grade level up) can be used to provide additional validity evidence.

The impact of the resulting cut scores on special populations (e.g., English learners, students with disabilities, migrant students) will need to be carefully considered when crafting the validity argument. Our design will show how these pieces of evidence can be used and combined to make a cohesive validity argument. This will help assure that SBAC standards correspond with national and international expectations for education.

**Collaboration**

The core philosophy that led to the partnership between ETS and Measured Progress is the common desire to build strong positive working relationships with our clients. These positive relationships result in collaborative partnerships between our organizations and our clients, where ideas can be shared and refined to develop and advance assessment programs across the country. By working with the SBAC, the SBAC TAC, PARCC, and the SSDC, we look forward to bringing our collaborative philosophy to fruition so that the resulting standard-setting design will draw from a variety of perspectives while meeting and exceeding the needs of the Consortium.
2.4 Determine Vertical Scale Design

A fundamental requirement of vertical scales is that content standards and assessments at different grade levels are articulated. This is consistent with the underlying developmental continuum inherent in the Common Core State Standards (National Governors Association Center, 2010) which refocuses instructional emphasis and facilitates inferences regarding change in academic achievement and readiness. Of course, and as the RFP indicates, until information is available regarding the dimensionality of each of the content area assessments, we feel it is unwise at this time to propose specific scaling approaches. We will work in collaboration with the Consortium and the TAC to specify the number and types of scales to develop in order to best reflect the underlying structure of the constructs, while simultaneously providing scores that are both understandable and useful to the Consortium and their stakeholders. We also expect to involve the TAC in design decisions and interpretation of results periodically, as determined by the Consortium. In the following section, we will describe our approach to investigating the impact of the construct dimensionality and our approach to constructing and evaluating a provisional scale with the pilot test data. We also include in this section, our proposal for establishing and evaluating the materials and procedures for scoring of the performance tasks.

Pilot Test

Data Collection Design and Sampling

Pilot Study Design. As described in Section 2.1, we propose a matrix data collection design that uses both common-item and equivalent groups links. These designs provide a framework that samples the test blueprint and links both horizontally and vertically, which will support all the pilot test studies and deliverables described throughout this proposal. In addition to the analyses described in Section 2.7, we will use the resulting data to evaluate the underlying dimensionality of the constructs and to evaluate the feasibility of creating a single common vertical scale, or multiple scales. We can also use the pilot test to evaluate the feasibility of linking both summative and performance tasks together in the construction of the vertical scale or scales. Although the Common Core State Standards (National Governors Association Center, 2010) have been developed based on an underlying developmental continuum, and therefore should support the development of a coherent vertical scale, this assumption requires empirical validation. Careful consideration of the evidence necessary to support the vertical scale or scales is a key element in the designs and methods described below. Our goal is to propose the most technically sound vertical scaling design then refine the design and concepts in consultation with the Consortium’s Executive Director, the lead psychometrician, and the TAC. Other design options that further reduce the burden of testing can be conceptualized as variations of those proposed here. In addition, there is the provision for the separate scaling of the summative tests and performance assessment components as directed by the Consortium and the TAC.
As initially described in **Section 2.1**, we are proposing two vertical matrix designs for the pilot stage. Each design employs both randomly equivalent groups and common items to link calibrations both within and across grade levels. Designs that incorporate both common-item and equivalent groups linking designs afford the Consortium the reasonable assurance that the data collected will yield sufficient information to link horizontally, as well as evaluate the potential vertical scale, or scales, that will be used in the field test and final operational assessment. In addition, these two designs are flexible and can be amended in collaboration with the Consortium and the TAC once the performance task demands are better known and the test blueprints are finalized.

**Pilot Test Design.** As described in **Section 2.5**, the pilot forms will be fixed linear forms administered online and will be assembled to represent the test blueprint. It will include all item types, including the performance tasks. For planning purposes, we assume that each student will receive a fixed linear forms of approximately 45 items in length and a set of 1-3 performance tasks. We are proposing to administer the form over approximately two hours of testing time, which could be divided into two sessions and scheduled across two days, paralleling the operational model where student are given the performance tasks several weeks in advance of the summative CAT. We recognize the test blueprints and specifications are not final, so will work with the Consortium in refining these designs and incorporating modifications and suggestions from the Consortium and the TAC.

As described in **Section 2.1**, to collect data for the linking study ETS will randomly assign students in each grade level to one of two groups: an on-grade sample or an cross-grade sample.

- **On-grade samples** — Students in this group will be administered forms that include items consistent with their grade level. The linear fixed forms will have some items in common across forms and will be administered alongside one or more performance tasks. These items will be used for horizontal equating studies, placing the on-grade level pilot test items on the same scale.

- **Cross-grade samples** — Students in this group will be administered forms that include:
  - Items consistent with their grade level (on-grade level items) from the pilot administration, as described in the previous bullet, and
  - Items from an adjacent lower, but not higher grade (cross-grade items), which will be used to validate, adjust and strengthen the vertical scale.
We also proposed two possible configurations for administering the forms. If you recall, in the first vertical scaling option, below-grade-level items and, if desired, performance tasks\(^1\), are administered to a single cross-grade student sample. Hence students in this cross-grade sample will respond to more items than comprise a single test form, and potentially many more items than in an operational adaptive administration. However, this design allows for strengthening of the vertical scale in situations where the performance assessments performed differently than other items, but were not so different that important separate dimensions were suggested.

In the second vertical scaling option, below-grade level items are administered to one cross-grade student sample, and below grade level performance tasks are administered to another cross-grade student sample. This option has the advantage of decreasing the testing burden for any given student, but it means that in the calibrations, some analyses will define the trait — including performance assessments and other analyses — will define the trait not including the performance assessments.

Either of these approaches will provide sufficient data to link the pilot test forms horizontally and to vertically. In addition, these two designs are flexible and can be amended as necessary in collaboration with the TAC, once the results of the pilot test are known and the performance task demands are better understood.

We understand that content and delivery platform concerns may place restrictions on the final configuration of the forms for the administrations, and will advocate on behalf of psychometric considerations (particularly the vertical and horizontal scaling design) to arrive at the best possible solution. ETS will work with the Consortium and its Contractors prior to the start of the pilot test administration to carefully identify the specifications that will result in the delivery of the horizontal and vertical linking items and will conform to the design requirements for administration and data collection. We will work efficiently and with the foresight that comes from long experience in partnering with organizations of all kinds in the successful attainment of assessment goals. We assume that this process will be done via email, conference calls, and videoconferencing.

**Sample.** As described in Section 2.1, we recommend using samples that are representative of geographic, demographic, and special needs groups using a stratified multi-stage probability sampling method. Our proposal is to administer the pilot forms to representative samples of approximately 1500 examinees per item/performance task plus, sufficient number of SWDs and ELs for the special analyses and the comparability study. We plan to take advantage of the flexibility inherent in the online test administration for spiraling the forms at the student level.

**Dimensionality Study**

\(^1\) As described under the pilot test section, we have provided a design that would allow for the scaling of the performance tasks with the summative tests. Should the Consortium not want to pursue this option, the performance tasks could be administered in a stand-alone field test design.
Vertical scaling makes the implicit assumption that the same essential construct is being measured at the top and bottom of the scale that reflects increased growth in the nature of the construct. Vertical scales provide a weaker level of comparability than equating, since somewhat different content is sampled across grades, which has different psychometric properties by definition. Yen (2007) suggests that vertical scales are like a folding ruler that curves through space when held out, where connections among some grades or clusters of grades are stronger while others are somewhat looser. This assumption of the same essential construct being measured may be more difficult to achieve when the vertical scale includes a wide span of grades. Caution is necessary in the comparison of growth from substantially-different parts of a vertical scale, due to possible shift in dimensionality. Inferences are typically strongest when limited to adjacent grade levels that are the most similar in terms of content and instruction.

When a major dimension is present in the data along with minor ones, this is referred to as essential unidimensionality (Stout, 1987). Essential unidimensionality relaxes the assumption of strict unidimensionality, and provides a justification for the use of unidimensional IRT models and vertical scaling. If essential unidimensionality can be established then substantial evidence will be supplied in support of using a unidimensional IRT Model and creating a single vertical scale. Given the proposed design, along with the inclusion of the innovative new item types, empirical evidence is necessary to understand the dimensionality of the Consortium assessment and the impact on test specifications, composite scores, and vertical scaling. We propose to complete the dimensionality analyses described in the next section to determine whether a unidimensional vertical scale can be supported, or if the results would be characterized better by a multidimensional vertical scale. Each dimensionality examination will sample content that reflects the full test blueprint articulated within and across-grade levels. Approaches adopted for assessing dimensionality will be determined in collaboration with the Consortium and the TAC.

**Methods for Assessing Dimensionality**

A number of exploratory approaches have been developed to examine the dimensional structure of a single test, although there is little consensus about which of the approaches is “best.” Reckase (2009) and Zeng (2010) argue for the use of three approaches: DIMTEST (Stout, 1987; Nandakumar & Stout, 1993), parallel analysis (Horn, 1965), and a vector approach (Reckase, Martineau, & Kim, 2000). As a first step, it is important to determine if the data are indeed multidimensional (with a minimum of two dimensions). DIMTEST is a nonparametric, IRT-based approach that examines the conditional covariance between two distinct, homogeneous subsets of items. In essence, it tests whether the reference composites for the two subsets of items point in the same direction at different points along the scale. If there is no significant difference between the dimensions measured by these items, the data can be treated as essentially unidimensional.

Once essential unidimensionality is ruled out, the next step is to identify the number of underlying dimensions. The two methods most commonly used for this purpose in practice are
principal components analysis (PCA) (Hotelling, 1933) and exploratory factor analysis; however, these approaches can be quite subjective and do not often identify the correct number of dimensions (Zeng, 2010). For instance, one of the issues with PCA is that it is difficult to differentiate minor dimensions from nuisance dimensions (i.e., the smaller components in a scree plot). Parallel analysis was developed as a way to address this issue by identifying the number of orthogonal components that are distinguishable from random noise. In this approach the eigenvalues from the item correlation matrix are compared against the eigenvalues for sets of uncorrelated, randomly generated data with the same number of items. Parallel analysis has been shown to be one of the best approaches for identifying the “correct” number of dimensions (Henson & Roberts, 2006; Zwick & Velicer, 1986), although a key limitation is that it performs poorly when the dimensions are moderately to strongly correlated (Zeng, 2010). This is particularly important, since the dimensions for educational assessments are almost always moderately to highly correlated.

To address the problem of identifying the appropriate number of correlated factors, Reckase et al. (2000) developed an extension of exploratory factor analysis that uses a vector representation of items — based on the factor loadings — and examines changes in angles between items as the number of modeled dimensions increases. In general, the average angle between vectors tends to increase when a larger number of dimensions are modeled, but once the “true” number of dimensions has been reached, modeling more dimensions only results in small changes. As such, the number of dimensions can be identified by determining when the average change in angles approaches zero with minimal variation. The angle between items can be computed as,

$$\alpha_{pq} = \arccos \frac{a'_p a_q}{\sqrt{a'_p a_p} \sqrt{a'_q a_q}}$$

where $a$ is a vector of factor loadings or item slopes for two distinct items $p$ and $q$. To evaluate the point where the average change in angles approaches zero, Martineau and Reckase (2006) used box-plots (or what they refer to as box-scree plots). Zeng (2010) examined the use of these box-scree plots as part of a simulation and found that they accurately identify the number of underlying dimensions, even when the dimensions are highly correlated. All of these approaches can be run using the open-source software R (R Development Core Team, 2011). While we believe that these approaches are defensible for assessing dimensionality within and across grades, alternative suggestions can be implemented at the discretion of the Executive Director, Lead Psychometrician, the Consortium, and the TAC.

**Detailed dimensionality procedures.** Traditionally, dimensionality examinations are conducted for tests at each grade level separately. If the tests are found to be essentially unidimensional for each grade, it is commonly assumed that the dimensional structure across grades is also unidimensional. This may not be the case. As such, it is necessary to examine the dimensional structure both within and across grades.
1) As a first step, we plan to use DIMTEST to determine whether the items at each grade level and for each pair of adjacent grades are essentially unidimensional. This will provide an initial justification for the creation of a unidimensional versus multidimensional vertical scale.

2) As a second step, we plan to examine the dimensional structure of the tests using both exploratory and confirmatory methods. For the exploratory approach we plan to use a parallel analysis and the vector approach developed by Reckase, Martineau, and Kim (2000). The purpose of this examination is to identify the number of dimensions at each grade level. For the confirmatory analysis we plan to examine the items using a between item dimensional structure (i.e., simple structure) where each item is only allowed to load on a single factor. The factors will be chosen to align with the Common Core State Standards. These results will be used to determine the extent to which the data departs from a unidimensional factor structure.

3) As a third step, a unidimensional and multidimensional vertical scale will be created. In the multidimensional case, the information from the second step will be used as the basis for the number of dimensions. Grade level means and standard deviations on the unidimensional and multidimensional scales will be compared to determine the extent to which the dimension-specific trajectories differ from the unidimensional trajectories.

4) The results of the dimensionality analyses will be reported to the TAC and in a technical report. These will include recommendations for the vertical scaling procedures to be used in SBAC.

**Evaluation of the Vertical Scale**

As described in Section 2.1, we propose to carefully evaluate the results prior to, during, and after IRT scaling to examine the integrity of the vertical scales. A fundamental aspect of evaluating a vertical scale is to investigate how examinee scores change over grade levels by assessing changes in means and standard deviations of scale scores, as well as changes in medians and selected percentile ranks. Prior to conducting the IRT vertical scaling, classical test statistics are examined to evaluate the ordinality of linking items. In general, we would expect to see higher performance for higher grade levels on the cross-grade common items.

As described in Section 2.1, we will evaluate the medians, means and standard deviations for all grade levels, frequency distributions, means, and standard deviation by school types and examinee demographic characteristics. In addition, during IRT scaling, we will monitor model-data fit and convergence criteria for each calibration run. After IRT scaling, we will compare the equated score distributions across grades to see if the vertical scaling results appear reasonable. We will examine the test characteristics curves (TCCs) across all levels of a content area to assess whether these curves match expectations given the particular content area, the way in
which the tests were constructed, and how students typically learn a particular content area. The TCCs for all forms on the new vertical scale, within a content area, will be plotted and examined for overall progression in difficulty. It is expected that, in general, forms will increase in difficulty as grade level increases. Likewise, the conditional standard errors (CSEMs) for all tests on a vertical scale will also be plotted and examined for overall ordinality. It is expected that as the test level increases, the ability level at which the test is measuring most accurately will increase as well. **Figures 3.a and 3.b in Section 2.1.1.C** show examples of how we would expect TCCs and CSEMs plots to look for a properly functioning vertical scale.

**Description of pilot output to be used in adjusting field test design**

Given that there is so little time from the completion of the field test and the launch of the operational program, the pilot test is pivotal to the overall success of the project. We will describe some of the outputs below, but remain open and flexible to incorporating additional outputs after consulting with the Executive Director and lead psychometrician for the Consortium. The results from the pilot will help inform the design of the field test in several ways, including: a.) evaluation item and task performance; b.) identify the impact of construct dimensionality; c.) construct and evaluate the IRT scale; d.) construct and evaluate a provisional vertical scale; e.) evaluate provisional paper test forms; f.) evaluate accommodated test forms; and g.) establish and evaluate materials and procedures for scoring the performance tasks. More detail is described in **Section 2.7** for each of these studies.

**a.) Evaluation of item and task performance.** Examining the content and psychometric characteristics of the items from the results of the pilot testing will aid in the evaluation of two key aspects of the instruments. First, we will evaluate item performance, as well as the distribution of items in terms of content and difficulty. Prior to the pilot phase, the psychometric characteristics of items — of particular importance the difficulty — were unknown. With knowledge of the pilot item statistics, items may be revised or administered at a different grade level during the field test stage to obtain updated item statistics. Items selected for use in the field test need to meet item and test selection criteria for classical and IRT statistics, as well as sampling the full test blueprint.

For any assessment, it is critical to have a solid representation of items that measure the full set of test specifications, and include item difficulty that represent the range of ability. This is particularly important in the development of an item pool for use in computer adaptive environments. If the goal is to improve measurement across the entire ability range, then the item pool must include items that measure the ability range with the proper content characteristics. The pilot test provides an opportunity to identify any remaining gaps in this content coverage, identify items that may need to be targeted to a different grade, and will determine the form assembly requirements for the field test. Results from the pilot test will also provide information about timing requirements, overall form difficulty, and estimated form reliability, and may also provide information to be incorporated into the field test sampling requirements. We will also be able to gather additional information about the performance of
innovative item types and be able to make revisions prior to the field test, if necessary. The pilot test will also provide an opportunity to evaluate and refine items and rubrics associated with both hand scoring and the automated scoring engines.

b.) Construct dimensionality. The results from the dimensionality studies described earlier will confirm essential dimensionality and provide initial justification for the use of a unidimensional IRT model and scale. The dimensionality studies will also determine the extent which the data depart from a unidimensional structure, and if they are sufficiently unidimensional to support scaling the CAT pool items with the performance tasks, or whether a composite score based on two separate scales is necessary. In addition, the results of these studies will be used to determine whether a single vertical scale or a multiple scale should be developed in each content area.

c.) Construct and evaluate the IRT scale. Described in detail in Section 2.7, the data from the pilot test will be used to select the IRT model that will be used for the Consortium’s operational adaptive test.

d.) Construct and evaluate a provisional vertical scale. The pilot test will provide us with information to evaluate the effectiveness of the horizontal and vertical linking designs. We will pay particular attention to the performance of the cross-grade common items between adjacent grades. In order to provide an acceptable vertical scale, the observed score performance on the cross-grade common items is expected to increase with each grade. Selecting cross-grade linking items that reflect growth in the construct and demonstrate good discrimination will help to stabilize and cross-validate the vertical scale. The expectation is that the validity of the resulting vertical scale will be further strengthened in the field test phase.

e.) Evaluate comparability of provisional paper test forms. The results of the comparability study described in Section 2.7 will be used to link the paper forms to the online forms. In addition, the results of the analyses will provide information to understand the effect of the delivery method and platform. We will use this information to make recommendations for changes and modifications in the finalization of the selection of items for the field test forms, future item development, and the delivery and administration methods of online and paper and pencil forms to minimize mode-effect differences in performance.

f.) Evaluate accommodated test forms. Using the results of the studies to evaluate the accommodated test forms, will work with the Consortium and the Technical Advisory Committee to incorporate accommodations and accessibility features to improve accessibility for ELs and SWDs and measurement of the construct. We will also use this information to develop a research agenda to further examine the effectiveness of the accommodations and accessibility features used in the Consortium assessments, and the degree to which they support fidelity to the focal constructs and provide valid assessment outcomes.
g.) Establish and evaluate materials and procedures for scoring the performance tasks. As the performance tasks will be an integral part of the summative score, maintaining consistency across prompts and raters is important to the integrity of the assessment program. Information from the studies described below and in Section 2.7 will be provided to the item development Contractor, the Consortium, and the Data Review committees. This information may be used to improve scoring process, training, or rater quality. In addition, the information may be used to improve the item development specifications, design patterns (Mislevy, Riconscente, & Rutstein, 2009), or task templates, thereby minimizing prompt specificity in the generation of performance tasks.

**Field Test**

The field test has two main purposes: 1.) calibrate all of the available items in preparation for the operational program; and 2.) validate, adjust and strengthen the horizontal and vertical scales established based on the pilot test data. What follows is the description of the design and analysis plans associated with validating the vertical scale. The data collection will occur in phase one of the field test, when linear fixed forms are administered. Please see Section 2.7 for complete details of the field test design and planned studies.

**Linking Design.** While the choice of a data collection and linking design is partly contingent on the results of the pilot test and will be determined in conjunction with the Consortium and the TAC; at this time we propose that we use the same hybrid model that is described in Section 2.1 and under the linking design for the pilot test (this section), incorporating the strength of the common-item and randomly equivalent groups linking designs. That is, assuming that we use a similar linking design, ETS will randomly assign students in each grade level to one of two groups: an on-grade sample or an cross-grade sample. Students in the on-grade group will be administered forms that include items consistent with their grade level. Students in the cross-grade group will be administered forms that include on-grade level items and items from an adjacent lower, but not higher, grade — which will be used to validate, adjust, and strengthen the vertical scale.

As with the pilot test, there are two possible configurations for administering the forms: a.) below-grade-level items and, if desired, performance tasks\(^2\), are administered to a single cross-grade student sample, or b.) below-grade-level items are administered to one cross-grade student sample, and below grade level performance tasks are administered to another cross-grade student sample. Either of these approaches will provide sufficient data to link the field test forms horizontally and to adjust, validate, and strengthen the vertical scale. In addition, these two designs are flexible and can be modified as necessary in collaboration with the TAC.

\(^2\) As described under the pilot test section, we have provided a design that would allow for the scaling of the performance tasks with the summative tests. Should the Consortium not want to pursue this option, the performance tasks could be administered in a stand-alone field test design.
once the results of the pilot test are known and the performance task demands are better understood. Please see Section 2.9 for details regarding the field test anchor item and task selection to support the vertical scaling study.

**Sampling.** In order to establish the quality and validity of the item parameters obtained during the field test, we propose to obtain approximately 1500 students, representative of the target population, by implementing a sampling plan for the first stage of the field test that is similar to that used for the pilot test. A stratified multi-stage probability sampling method was proposed to select participating schools within each state, with oversampling of SWDs and ELs for special studies and scaling of special forms. Additional variables of interest defined in collaboration with the Consortium and the TAC would be used in identifying the schools for recruitment. For the adaptive stage of the field test, a similar sampling plan could be implemented to invite participation among those schools with the appropriate technology. However, if many of the schools in the Consortium want to participate in this phase of the field test, we can sample within the resulting participants to achieve a representative sample.

**Scaling and Computation Procedures.** In this section, we first describe the calibration of new and revised items against the anchors. We then discuss evaluation of the vertical scale, followed by the calibration and linking of the new items to the vertical scale.

a.) **Item calibration against anchor items.** Following receipt of student response data from the fixed form online administration, items and tasks will undergo classical test theory, differential item functioning, and item response theory analyses as described for the pilot in Section 2.7. Item/task statistics will be evaluated against approved criteria and flagged items sent for review by the Consortium and its test development Contractor. All items determined to have satisfactory psychometric properties will then be linked to the vertical scale by horizontal equating to their respective grade scales using the embedded pilot anchor items. Anchor items with large changes in item parameter estimates between pilot and field test administrations will be subject to review and possible removal from the anchor set. The approach used for horizontal linking will be that approved by the Consortium and the TAC.

b.) **Evaluation of the vertical scale.** The basic approach is to cross-validate the vertical established in the pilot. The field test design described above will provide a rich set of items to be used for the evaluating the vertical scale and strengthening it as needed the vertical scale. The vertical scale established during the pilot will be evaluated with respect to the ordinality of IRT test characteristic curves across grade, and stability of between grade relationships. The same analyses, described in Section 2.1.1 and referenced earlier in this section under Pilot Test, will be use to evaluate the vertical scale.
**Evaluating the Performance Task System**

As the performance tasks will be an integral part of the summative score, it’s important to establish that the results are consistent across raters and prompts. This can be accomplished by evaluating the results of the administration and through careful management of the scoring processes, minimizing all possible sources of variance associated with the scoring process. In order to minimize any sources of irrelevant variance, we propose a comprehensive set of plans for evaluating and monitoring the scoring systems in the pre-operational stage through the operational stage. The studies described below will provide a basis for monitoring whether the score categories and the underlying construct is maintained in the field test and subsequent administrations. In addition, information from these analyses will be provided the item development Contractor and the Data Review committees for review, for suggestions to improve the item development specifications, design patterns (Mislevy, Riconscente, & Rutstein, 2009), or task templates, thereby minimizing prompt specificity in the generation of performance tasks.

**Pre-Operational.** Pre-operational procedures consist of range-finding, selection of calibration/benchmark papers, and the establishing materials for rater training and qualification. Well-developed process and procedures in the pre-operational phase determine the success of the operational phase.

a.) **Certification and training.** We recommend that following the successful completion of a certification test, each qualified rater receives rigorous training in correctly applying the rubric at each specific score point. These should be documented in the Technical Report. Both during training and throughout the entire scoring process, raters should be guided by the benchmark and rangefinder responses selected during the range-finding process, which serve as the models of the level of quality required for each score point. We also recommend that after training is complete, raters take daily calibration tests to further establish that they are scoring reliably and consistently. They should be monitored during scoring by scoring leaders who constantly review and analyze raters’ scores using interspersed validity responses that provide ongoing measures of rater accuracy. If the scoring Contractor has this capability, we will regularly analyze real-time data, such as those collected electronically by online scoring systems, to further establish the reliability of the scoring process.

b.) **Automated scoring.** If automated scoring is used, then the engines will also have to be set up and “trained” to score the prompts. We expect that the majority of this work will be completed by the automated scoring Contractor. We expect to be involved in the set up and monitoring of the system to establish that the engines are scoring consistently. In order to establish that the items are scored consistently by the scoring engines when the engine is updated to a new version, we recommend that a special study be completed to evaluate the new version’s impact on the resulting scores.
c.) **Calibration papers.** Calibration papers with known psychometric properties are selected by experts and establish the standard for scoring various types of responses — these are also known as “benchmark” papers. These papers are distributed periodically during the course of scoring and are critical to determining the accuracy of scoring. In the pre-operational phase it is critical that large and robust calibration papers are able be selected; papers that represent the expected types of student responding. Documentation of the procedures used to select calibration papers and the number and distribution is important to the accuracy of scoring. Calibration papers are a property of an item that is important if the task is administered in the future and consistency in score assignment is necessary.

d.) **Range-finding validation.** We assume that the proposed rubrics and range-finding responses will be identified by the scoring Contractor, and recommend that the proposed rubrics be applied by at least two judges to at least 200-300 responses per subject and grade level. The actual number that is required will depend on the form of the performance tasks, characteristics of the automated scoring model, and characteristics of examinee responses to the tasks. We will ask that these judges rate responses to all performance tasks within the form being evaluated, and will request that the scoring Contractor provide the resulting data to us. During the validation of the range-finding, we assume that the content expert panel will review the scoring rubrics, raters’ abilities to consistently apply these criteria, and the ability of the automated scoring model to match human raters’ scoring. We assume that calibration/benchmark papers for any tasks that require ongoing rater scoring will also be identified during these meetings. Based upon the values of these agreement indices, we will work with the Consortium to establish flags to identify items that are not performing consistently with expectations, requiring careful review in evaluating whether automated scoring is appropriate for the item/task. We will provide the full complement of rater agreement statistics for use in a meeting with the scoring Contractor and the external content experts to finalize the performance task rubrics, and to identify benchmark student performances.

**Operational.** Operational scoring procedures include monitoring the method of distributing student responses, as well as real-time monitoring of rater accuracy and consistency and supervisory review and auditing. These studies are described below. We will work with the Consortium and the scoring Contractor to support these studies and conduct the proposed analyses:

a.) **Distribution of responses.** The method of collecting student responses and distributing these to raters needs to be consistent in order to prevent introducing irrelevant variance. To verify that images that reflect the properties of a students’ response to be scored are rendered accurately, we propose to conduct an audit of the process and system and collect samples of the student work in collaboration with the scoring Contractor. These will be done periodically throughout the scoring window. In addition,
the consistency in score assignment over different jurisdictions will also be examined. We will include a description of how optical images of student responses are distributed to raters in technical documentation and the results of the audit activities.

b.) Ongoing training and calibration. For those performance tasks that require rater scoring, we propose to address consistency of human ratings during the scoring window. Early in each scoring session, we will ask the scoring Contractor to have each rater score 8 to 10 monitor/calibration/benchmark papers. Several indices will assess the extent to which a rater is well-calibrated to score. These include the percent of exact agreement, the sum of differences between each rater’s scores and the consensus scores, and the sum of squared differences. All of these indices will be evaluated to determine if remedial action should be taken.

c.) Rater monitoring within Occasion

i.) Human constructed response scoring. The statistics and methods used for monitoring rater agreement are more thoroughly described in detail in Section 2.7. Statistics produced for evaluating the functioning of performance tasks may include, but are not limited to:

» Number and proportion of students earning each rubric score;

» Number and percentage of exact agreement between two human ratings or between automated and human scores after correcting for chance agreement rates;

» Number and percentage of adjacent agreement between two human ratings or between automated and human scores after correcting for chance agreement rates;

» Number and percentage of non-adjacent scores between two human ratings or between automated and human scores after correcting for chance agreement rates;

» Unweighted Kappa statistics (Cohen, 1960), which characterize the degree of agreement or association between two human ratings or between automated and human scores after correcting for chance agreement rates;

» Quadratic-weighted Kappa statistics, which have similar properties to unweighted Kappa, but are degraded disproportionately by the presence of large disagreements between ratings of two human raters or between human and automated scores; and

» Pearson correlations, which provide another measure of the degree of agreement or association between two human ratings or between automated and human scores.
Automated scoring. Some validation of scoring even for automated algorithms is also necessary. Consistent with the procedures used with rater protocols for monitoring reliability, the following statistics can be produced:

- Similarity of human and automated score frequency distributions and moments (means and standard deviations).
- Standardized differences (effect sizes) between human and automated score means. This is computed as the difference between means divided by the standard deviation of the human scores.
- Unweighted Kappa statistics (Cohen, 1960), which characterize the degree of agreement or association between automated and human scores after correcting for chance agreement rates.
- Quadratic-weighted Kappa statistics, which have similar properties to unweighted Kappa, but are degraded disproportionately by the presence of large disagreements between some human and automated scores.
- Pearson correlations, which provide another measure of the degree of agreement or association between automated and human scores.

d.) Rater agreement and accuracy across occasions. There are three possible ways that performance tasks might be scored. The approach we propose to evaluate that scoring across occasions is different for each scenario.

i.) If an automated scoring model is applied to all performance tasks in a particular family, we recommend that the scoring vendor enlist human raters to score a small number of responses at each administration to confirm that the automatic scoring is functioning as expected. The absolute number of such ratings will depend on the form of the performance tasks, characteristics of the automated scoring model, and characteristics of examinee responses to the tasks.

If a difference is observed between the automated and human scoring, we recommend that the scoring Contractor have the same group of raters assign scores to responses from a historical administration. Consistency with historical human ratings will suggest that new human ratings have not changed, and the automated scoring would be suspect for the particular prompt at issue. Inconsistency of new form human ratings with historical human ratings might suggest a training problem, and would therefore suggest that the raters be recertified and the papers be scored again.

ii.) If the automated scoring approach implemented requires that a new scoring model be trained for each new task, then human scored responses will already have been collected in order to support the training and evaluation of those models for each item. This human-scored response data is sufficient to establish the reliability of scoring, but does not obviate the need for human scoring of a
sample of responses in each operational administration where the item is used, so that discrepancies in item response characteristics can be tracked. The procedure used for resolving inconsistencies between automated and human scoring in the previous case would also apply here.

iii.) **Rater trend study.** If human scoring of a task or tasks occurs across occasions (or administrations), a more in-depth evaluation would be recommended called a trend study. Tate (1999) articulated a solution to the problem of subjective or changing scoring standards for constructed-response items (similar to the procedures followed in Fitzpatrick, Ercikan, Yen, & Ferrara, 1998). He suggested an equating study in which any across-year changes in rater severity could be isolated, so that across-group ability differences could be adjusted, if necessary. The equating study envisioned by Tate involves rescoring responses to the constructed-response items obtained from a previous administration. These responses, obtained from the previous group of examinees, would be scored by the same raters scoring responses for the same items for the new group of examinees. These trend papers have two sets of scores associated with them: one from the old set of raters and one from the new raters, which are then compared.

If this approach is adopted by the Consortium, we will designate one testing cycle as the baseline cycle. Responses from this baseline cycle will be randomly selected and scored during each future cycle. The scores from the future cycle will be compared to the baseline scores to obtain a conversion table of equated scores. Any of several criteria could be used to determine whether scoring standards have changed sufficiently from the baseline year so that it is necessary to use the converted scores. We recommend the use of the conditional standard error of equating (CSEE) from the mean-standard deviation linear equating approach. An asymptotic approximation to the CSEE is given by,

$$CSEE \approx \sqrt{\frac{s_Y^2(1 - r_{XY})}{N}} \left\{ 2 + (1 + r_{XY}) \left[ \frac{X_i - \bar{X}}{s_X} \right]^2 \right\}$$

where $Y$ represents the set of baseline scores; $X$ represents the new set of scores; $\bar{X}$ and $s$ are the mean and standard deviation, respectively, of the scores; $r$ is the correlation between the two sets of scores, and $X_i$ is one of the possible score levels for the assessment. If the original score and the converted score differ from each other by more than twice the CSEE at any point, this will be taken as evidence that a non-negligible shift in scoring standards has taken place.

A second optional step consists of new set of ratings being directly equated to the old set using the mean-standard deviation linear equating approach. The outcome is a table listing, for each score from the new raters, the corresponding score if the previous
raters had scored the entries (i.e., the equated scores). All scores for the current year are converted to scores from the previous year using this table.

Even if trend scoring is not used, we will summarize the responses for each performance task after scoring is complete. These summaries will include statistics such as:

- Frequency distribution of raw task scores,
- Frequency distribution of equated multiple choice scores, and
- Correlation between task scores and multiple choice scores.

**Generalizability Study.** In addition to the omnibus rater-agreement statistics presented above, we will conduct a generalizability study to investigate the contribution of different facets, such as raters, students, and tasks to overall variance. We will also examine consistency in score assignment over different jurisdictions (sites). We will conduct a generalizability study for selected grades and content areas to establish that scoring site or occasion is not contributing unwanted variance. The proposed generalizability studies will analyze the raters’ evaluations of no fewer than 2000 student responses for each set of constructed response items for each grade and content area. We will use this analysis to partition the score variance into its components and estimate the variance attributable to each facet in the design. The inclusion of the rater facet into the design will allow us to estimate the rater variance and the variance associated with interaction between the rater facet and other design facets. Following the generalizability study, we will conduct decision studies to estimate the generalizability and dependability coefficients appropriate for the given test design. In this study, we will investigate the impact of different configurations of task types, as well as the impact of number of raters and rater reliability, on the generalizability of SBAC scores. Please see Section 2.7 for a detailed discussion of the generalizability study.

**2.5 Design Pilot Test Sampling Plan and Select Items and Tasks for Pilot Test Forms**

Although the test blueprints for the Consortium’s assessment are not currently available, the goals and design of the pilot, in combination with the constraints for the pilot forms stated in the RFP, provide a framework that we can use to describe our proposed approach for pilot form construction. To meet these goals and timelines of the new assessment system, pilot forms for each grade must be developed to:

- Follow the operational test event blueprint, as pertains to representation of the specified content coverage as entailed in the Common Core of Standards, depth of knowledge, and designated mixture of item types;
- Target the difficulty level designated for each grade level, and include items that span a wide range of difficulty and depth of knowledge (DOK);
» Support within-grade (horizontal) and cross-grade (vertical) linking for development of the preliminary vertical scales that reflect growth in the construct;

» Yield a sufficient number of items to serve as anchors during the field test;

» Yield a sufficient number of additional items for the interim assessment item pool;

» Support the development of three years of fixed form versions of the adaptive test design for paper and pencil tests for students, without access to electronic platforms and special needs forms; and

» Select innovative items types that may be scored automatically.

As noted in the RFP, the pilot forms will be subject to the following constraints:

» Fixed form, administered online but not computer adaptive;

» Include all item types (i.e., selected response, constructed response, and technology enhanced);

» Online and paper/pencil forms; and

» Pilot testing also includes pilots of performance tasks. These tasks will be given to students and rated using operational rubrics and rating processes.

To begin item selection for the pilot test, we will study the Consortium’s test specifications and blueprints for each content and grade to fully understand the expectations for how the developmental progression toward college and career readiness articulated in the Common Core State Standards will be operationalized in summative assessments and performance tasks for each grade level. Based on this review and in collaboration with the Consortium, its technical advisors, and content experts, we will develop specifications for the vertical linking items, and for the selection of on-grade items to populate the pilot test forms. Preliminary forms assembly guidelines will be developed for item selection to fill the requirements of the test blueprints and specifications, including those of the selected linking design. This is an iterative process in collaboration with all parties identified by the Consortium to finalize the test assembly guidelines.

Using the established guidelines, content experts — chosen by the Consortium — will select items representing specified content coverage, depth of knowledge, and item types as well as the identified range of difficulty. Because the statistical properties of the pilot items will be largely unknown prior to pilot testing, the selection of items to populate the pilot forms will require expert judgment in relation to difficulty and expected performance of the items, and their suitability for inclusion in the on-grade pilot form and in the vertical linking sets. These judgments may be based on the alignment of items to the Common Core State Standards and the articulation of content across grades. In the absence of data, evidence centered design
(Mislevy, Steinberg, & Almond, 2003) may help identify item difficulty drivers. A new approach called design patterns could also be utilized. Design patterns (Mislevy, Riconscente, & Rutstein, 2009) consists of attributes that are associated with components of an assessment that correspond to identifying the knowledge, skills, or abilities that test developers want to capture. Very small classroom pilots will provide preliminary information on item characteristics. We will provide consultation concerning any psychometric implications, as necessary.

In the following pages, we provide a general description of the process we typically follow for selecting items for linking, for fulfilling the test blueprint, for the development of the field test anchors, and for populating the interim assessment item bank.

**Selection of Cross-Grade Linking (Anchor) Items**

An appropriate set of cross-grade linking items is required to establish the Consortium’s vertical scale, and as such it requires careful consideration from both the content and psychometric perspectives. Given the importance of the development and maintenance of the vertical scale, we propose that the cross-grade linking items will be selected first. For more details for the linking plan, please see Section 2.1.1, Pilot Designs. One step concerning the vertical scale linking design involves assigning students randomly to take either cross-grade or on-grade items. For the cross-grade items, some consideration as to how those items will function is necessary that entails hypotheses about how content articulates across grade levels.

We anticipate that this will be an iterative procedure, requiring close collaboration with the Consortium’s content experts. As cross-grade linking items will be delivered to an on-grade student sample (say at grade 3) and a cross-grade student sample (say at grade 4) the cross-grade linking items must first be appropriate for students in the lower of the two grades (in this case, grade 3 students). This will serve as the comparative basis for judging item performance when presented in the cross-grade pilot form (grade 4). To this end, we will work with the content experts to select a set of items that represent a range of difficulty (as judged by the content experts). This will help establish that when the vertical linking item sets are delivered to the cross-grade sample (grade 4), the items are not too easy for this group of students; we want to establish that the data are appropriate for item calibrations, and setting the scale if the items are too easy then the data may not yield usable results.

Additionally, vertical linking items selection needs to take into account that not all content standards are applicable to all grade levels. This is different from horizontal equating, in which the same content is represented on each test form. We will pay careful attention to establish that content experts identify standards that are applicable to both the on-grade and cross-grade samples of students. We will work with the content experts to help establish that the cross-grade link sets are representative of the constraints as specified in the RFP. After thorough internal reviews and approvals, we will submit the proposed linking sets to the Consortium for review and approval.
Some other issues affect the selection of items targeted at paper administrations. To the extent that innovative items are present that cannot be transferred to paper administrations, this will likely be a source of non-comparability between modes. This will need further discussion once the properties of these items are better known. Several online forms will have the same items as the paper administrations.

**Selection of Items to Fulfill the Test Blueprint, and for the Development of Field Test Anchors and the Interim Item Banks**

Upon approval of the cross-grade linking items, we will work with the Consortium’s content experts to select proposed sets of items to meet the specifications in the test blueprint for each grade that are consistent with the sampling plans as laid out in Sections 2.1 and 2.4. Given the need to populate the field test anchor pool and the interim item banks, the number of pilot forms to be developed will be based on the degree of overlap between forms, and the expectations for survival rates. These items selected for each form must reflect the operational test event blueprints, and the resulting set of operational items must adhere to all requirements of the test blueprint and test specifications. This item selection process involves attention to many factors. We will work in supporting the content experts to:

» Select the target number of items by reporting category or sub-category;

» Select a variety of item types as specified by the test blueprints and specifications;

» Select as broad a range of topics or reading passages as possible within each reporting category or sub-category;

» Select different types of items to measure a given standard to the degree possible;

» Select a range of items based on expected difficulty in accordance with test specifications;

» Select a range of items based on cognitive complexity in accordance with test specifications;

» Consider as even a distribution of keys as possible;

» Avoid items that may clue the answer to another item;

» Confirm that all selected items have appropriate expected item difficulty, and have been reviewed for bias and sensitivity; and

» Make some estimates as to expected testing time by form.

Following the selection of these items, the last step is to consider the sequencing of all items within the pilot forms.
Review of Pilot Test forms. Following final selection and sequencing of items composing the pilot forms and the cross-grade linking sets, we will send the proposed pilot forms to the Consortium for review. At this stage, revisions may be required. We will work collaboratively with the Consortium and their content experts throughout this process until pilot forms are finalized.

Pilot Testing of Performance Tasks

As noted in the RFP, we will include performance tasks in the pilot, and we will rate them using operational rubrics and rating processes or by automated scoring. We proposed a pilot design for performance tasks in Section 2.1.1, and briefly described the data required and associated analyses to evaluate task performance and the rating system, for which more detail is provided in Section 2.7. In this section we discuss how we will use these analyses to determine whether we need to adjust tasks or rubrics.

We will develop operational rubrics and use them to score responses during the performance task pilot. Following the scoring, we will analyze student responses and scoring data to determine if the tasks themselves — and the scoring rubrics — have performed as intended, or if they need to be revised or adjusted.

As described in Section 2.7, we will conduct a comprehensive set of analyses to evaluate pilot task performance and scoring. These analyses include classical, differential item functioning, item response theory, rater agreement statistics, and generalizability analyses. The results of these analyses will inform the main question that we would expect to answer on the basis of pilot testing of performance tasks — that is, if each task is appropriate for the purpose and design of the assessment and for the population of students being assessed. If the analysis of responses and data indicate that a task is not performing as expected, the Consortium may wish to exclude the task from the item pool. If, however, the task requires only surface revisions and adjustments in order to function well for operational testing, the Consortium may wish to evaluate the task with focus on the following features:

- Task directions—are these clear, precise, and appropriate for the task?
- Task wording—is the task phrased clearly enough, or do test takers seem confused?

If either the wording or directions is deemed to be problematic in a task that has otherwise performed well (as determined by statistical analyses), adjustments may be made and the performance task can be re-administered during field testing or rescored during the pilot if the defects are detected early enough and are amendable to adjustment in scoring.

Another important function of a pilot test of performance tasks is to confirm that the number of score points on the operational rubric is appropriate and yields meaningful distinctions in the quality of the responses. If, upon analysis of the responses and scoring data, the content experts
and psychometricians find that the responses exhibited a wider range of ability levels than score points available in the rubric, we may need to expand the score scale. If, on the other hand, the responses displayed fewer distinctions or produced fewer distinct ability levels than were anticipated in the original design of the rubric, we may need to contract the score scale.

We may need to adjust the rubric criteria themselves after pilot testing, and will make this determination on the basis of an analysis of the responses and the scoring data. The pilot test, for example, may reveal that the scoring criteria are not precisely aligned with the task or the task directions. In such cases, scrutiny of the responses may show that some features of the responses have been either inappropriately rewarded or inappropriately penalized, prompting the test developers to examine, revise, and clarify the criteria before the rubric is used operationally.

Following statistical analysis of the performance task pilot data, we will work with the Consortium and its content experts to review the results, and identify tasks and/or rubrics that may need revision. In the event that automated scoring is implemented for some tasks, we will collaborate with the scoring and test development Contractors to deal with any additional adjustments or analyses required.

2.6 Develop Pilot Test Item and Task Data Review Materials

After pilot testing, it will be important to involve educators from across the Consortium in a thorough review of all pilot test items and performance tasks, including information on the performance of items under different conditions (e.g., paper and pencil and translated versions). An examination of the actual statistical results from the pilot test for all items and performance tasks will be an important part of this review. Clear communication and presentation of information is paramount to a successful data review meeting, but will be even more critical given the introduction of new item types, new delivery methods and content associated with the new CCSS. The data and information must be presented in a clear and concise manner to maximize the opportunity for educators to discuss the observed results and provide feedback that can be used to prepare items for field-testing. In the case of the performance task components, there will need to be additional training for the educators attending the meetings on evaluating the data from these items and the scoring methods and systems as this may be unfamiliar to some attendees. As noted in the Q&A, due to the large volume of items, the data review meetings may be conducted online, through a distributed system, with face-to-face meetings limited, possibly to training or general orientation. We have extensive experience conducted both face-to-face and virtual meetings, and look forward to developing a system for data review that will meet the Consortium’s needs for educator engagement and efficiency in the review process. Our approach to preparing for these reviews are described below and we expect to work closely with the Executive Director and Lead Psychometrician of the Consortium in planning and designing these meetings to maximize our opportunity to share and gather information with educators.
Step 1. Establish Criteria

Prior to the Data Review meetings we will review the literature, existing best practice documents and gather information from Consortium members to provide the Consortium staff with some preliminary item and task evaluation criteria for consideration. We will then meet, and together we will establish the criteria that will be used in preparation and during the data review meetings for classifying item statistics. At a minimum we expect that these criteria will include item flagging, individually or in combination, to signal additional scrutiny for content appropriateness, bias, sensitivity, and overall statistical performance relative to expectation. Some examples of commonly used flagging criteria include:

- p-values above or below a specified threshold (e.g., above 0.95 or below 0.25)
- Item-total correlations below a specified threshold (e.g., 0.20)
- Distractor-total correlations above a specified threshold (e.g., 0.00)
- A greater number of high-ability test takers choose a distractor than the keyed response (generally students whose scores are at the 80th percentile or higher are used to define this high-ability group)
- A high percentage of omits (e.g., greater than 5 percent)
- A high percentage that do not reach the item (e.g., greater than 5 percent)
- IRT b-parameter estimates above 3 or below -3
- IRT a-parameter estimates less than .20 or greater than 2.0
- Items are also be flagged for IRT model-data misfit. A number of fit statistics are available, and choice of scaling model may influence the choice of fit indices. For example, if the Generalized partial credit model (Muraki, 1992) is used, fit statistics of interest may include Q1 (Yen, 1981; Fitzpatrick et al., 1996), G^2 (McKinley & Mills, 1985; Mislevy & Bock, 1990), S-X^2 or S-G^2 (Orlando & Thissen, 2000; Orlando & Thissen, 2003), and X^2 or G^3 (Stone, 2000; Stone, 2003; Stone & Zhang, 2003). If Masters’ (1982) partial credit model is used, fit statistics typically used are infit and outfit measures (Wright & Masters, 1982).

Once the criteria have been determined, we will then work with the Consortium on developing the report format, supporting documentation, as well as a training and communication plan that will be made available to all members of the consortium prior to the data review meetings. The format of the materials and report to be used in the data review meetings will be developed in consultation with the Consortium, however we expect that at a minimum the report will include the results from the analysis as well as any recommendations for improving the items or tasks. In order to maximize the time available to work with the educators, we propose to provide all attendees with briefing documents that describe the criteria, the rationale for the criteria, and
exemplars prior to the meeting. At the beginning of the meeting these materials will be reviewed and all questions and concerns answered clarified prior to the start of the meeting.

**Step 2. Conduct Analysis and Develop a Report**

After the analysis described in Section 2.7 are completed, the data will be reviewed and evaluated. We propose to review these results and the summary of item and task performance based on the criteria established in Step 1 and then to brief the Consortium and the item development Contractor in order to obtain initial feedback and to get approval to proceed with any revisions or additions to either the criteria or the set of analysis. After the preliminary review, we will incorporate the feedback from the Consortium and the item development Contractor and prepare the reports and supporting materials describing the results and any recommendations for items and task writing in preparation of the data review meeting. After final review and approval by the Consortium, we will prepare the documents for the data review meeting, send the briefing materials to the attendees, and post them for all Consortium members.

**Step 3. Attend Data Review Meetings**

With the briefing documents and data review materials finalized and produced, our psychometrician and program manager will then attend the two day committee meeting to be held in 2013, and work with the item development Contractor to guide participants through the data and preliminary recommendations. In order to get the most effective feedback from the educators, it will be important to provide time in the agenda to brief them on the process and answer any questions that they may have and clarify terminology, criteria, and expectations.

While the exact order of the agenda is subject to discussion and approval by the Consortium, we propose to start the meeting with a brief training session targeted to the audience of educators on how to interpret and use the statistics in the materials to review items. This would then be followed by a session conducted by our psychometricians, in collaboration with the item development Contractor and the Consortium, reviewing the briefing documents and data review materials, clarifying any concerns, and answering any questions. Once the committee members understand the materials and their task, we will then present committee members with data on all items, instructing them to carefully scrutinize any item or performance task that is flagged for additional review based on the established criteria. The committees will be asked to identify possible reasons for the results — whether they have detected a content problem within the item or task or whether there are instructional issues that may have negatively affected student performance. The committee will also be asked to recommend whether each flagged item and performance task is to be marked for revision, rejected, or accepted for inclusion in the Consortium item bank.
In addition to the review of the item statistics, the committee members will also be asked to
determine content appropriateness in light of empirical results. In addition to the standard item
statistics, for the performance tasks, we will also be providing information on task inter-rater
reliability (IRR), prompt comparability, and overall scoring reliability (this is described in more
detail in Section 2.7). We will ask committee members for their feedback as this information will
be important to consider for any potential revisions. For example, tasks with low reliability may
require adjustments to the task content and/or scoring rubric. Information on the analyses to
obtain this information is provided in Section 2.7. In addition, we will present information about
the statistical performance of remote (distributed) scoring systems for the committees’ and
Consortium’s information. More information on the analyses to obtain this information is
provided in Section 2.7.

In addition to review by committee members, any item or performance task that shows a
difference in item function between subgroups will also be reviewed by assessment specialists
from the item development Contractor and representatives of the Consortium. If an item or
performance task is found to contain material or information that appears to give an unfair
advantage to any group of students, or has bias and sensitivity issues, it will be revised for or
rejected from future use.

As noted earlier, it may be necessary to conduct data review meetings in a variety of formats,
with face-to-face meetings limited to training or general sessions. Depending on the volume of
items, the Consortium may also wish to consider online orientation and/or training sessions.

We would anticipate that during the early stages of data review there will be a greater need for
face-to-face meetings, but as the pool of experienced reviewers increases, orientation meetings
may move to online formats and greater use of distributed data review.

**Step 4. Reconcile Feedback for Final Decisions Working with CONSORTUM
and Report Final Needs for Item/Task Editing based on Pilot Review to
Prepare Items for Field Testing**

ETS will collaborate with the data review facilitators to document all educator feedback that
arises during the course of the meetings. After the data review meetings, we will review all
committee recommendations and any issues that were discovered during the review. In
particular, we will note items where the committee had difficulty reaching consensus on
recommendations for revision, rejection, or acceptance for inclusion in the Consortium’s item
bank, and the source of the difficulty. We will work with the Consortium to reconcile this
feedback and make final decisions for all items and tasks. We will prepare a report outlining final
needs for item and task editing based on pilot review to prepare items for field testing.

**Step 5. Finalize and Upload Item Statistics**
After completing the analysis, and after resolution of the status of every item, all statistical information, annotations and educator feedback, and the final statistical status will be uploaded into the Consortium item bank using the identified database layout provided by the Consortium. We will include all of the required variables within defined fields in the same upload file. The data format to be used has not yet been determined, but we expect to begin with industry standard approaches as we work with the item bank Contractor to produce the final format. We are experienced at providing data in multiple formats depending upon client needs or requirements.

Formats may vary in order to accommodate the specific test designs, but the general layout is fairly common. For example, we have experience with producing custom file formats when delivering information earmarked for external/non-ETS item banking systems. In most cases, file formats are derived from industry standard formats ranging from fixed length data records to QTI XML. A few of the file types we support at ETS include: fixed length record, comma delimited (i.e., CSV, XLS, XML), and QTI XML. Regardless of the final agreed-upon file format, ETS is flexible enough to adapt to meet the input requirements of the Consortium item bank. Once the contract is awarded, we expect to work closely with the Consortium and its item bank Contractor to understand the requirements and specifics of the item bank statistical upload process. At that time we can determine the best approach and anchor the data exchange format.

**Step 6. Produce Final Report**

The final task associated with the data review meetings for the items and performance tasks is to create a final report that will outline the recommendations from the committee members for managing the reviewed items. The report will reflect the final decision for addressing the flagged items and other observations from the committee, item development Contractor, our psychometricians and the Consortium. Items may be identified for revision and re-administration or for removal from the pool. Still other flagged items may be designated as having no revision but must be administered again as a field test item in a future form if the committee members believe the items is not problematic, but there was some other reason for the observed item statistics (e.g., motivation, opportunity to learn). For items designated for revision, the report will include the specific recommendations provided by the committee members. Any guidelines or suggestions for revisions to the item development process will also be captured and included in the report. For the performance tasks, information on scoring agreement and reliability overall and during the scoring window, as well as the statistical performance of the remote (distributed) scoring system, will be included.
2.7 Conduct Psychometric Analysis to Support Pilot Test Data Review

The data collection design for the pilot test that we have proposed and described in Section 2.1 will support several studies including:

1. Evaluating item and performance tasks,
2. IRT model selection,
3. Test form information,
4. Evaluation and monitoring of performance task scoring,
5. Accessibility for students with disabilities and English learners, and
6. Comparability of paper and pencil and online versions,

In the following section we describe our proposed plan to complete each of these studies, along with the available methods and software that is available to use. The results of these analyses will be provided the Consortium, used in the pilot data review meeting, summarized and described in the Technical Manuals and, where appropriate, loaded into the Consortium item bank.

1. Evaluating item and performance tasks

We propose to conduct a complete series of analyses that are designed to evaluate the performance of test items, tasks, and forms. We propose to complete the following analyses and will incorporate additional analyses at the suggestion of the Consortium and the TAC. In this section we describe the classical item analysis, differential item functioning (DIF) studies, and discuss the consideration and studies associated with selecting the IRT model.

A. Classical Item Analysis

The first step that will be completed will be a classical item analysis, which will be used to evaluate item difficulty, discrimination, and total raw score performance. We use these analyses to identify any items that might not perform as expected and to determine the score reliability of the summative tests and performance tasks. We will flag items that do not meet the psychometric criteria for acceptance decided upon by the Consortium. Our psychometric staff will work with the assessment development Contractor to carefully review each of these items to determine the degree to which they are appropriate for use in the operational tests.

We will conduct classical item analyses that include the following:

1. Classical item difficulty (or p-value) – For dichotomously scored items, this statistic indicates the proportion of examinees in the sample that answered the item correctly. Desired p-values generally fall within the range of 0.25 to 0.95. For polytomously scored items, this statistic represents the average item score or the proportion of the maximum obtainable score. Desired values generally fall within the range of 30 percent to 80 percent of the maximum obtainable score. Items that fall outside the desired
difficulty ranges can be justified for inclusion in an item bank based upon the quality and importance of the item content, or the need to measure students with very high or very low achievement levels. Items not functioning at one grade level might function better targeted at another one.

2. The percentage of students choosing each response option (selected-response items) – These statistics indicate the percentage of examinees that selected each of the answer options and the percentage that omitted the item. Properly functioning answer options are essential to each item. An item option not selected by any students indicates problems with the plausibility of that option.

3. Item-total correlation – This statistic describes the relationship between performance on the specific item and performance on the total test; sometimes referred to as a discrimination index. Typically, values of 0.20 or higher indicate a desired positive relationship with the targeted construct. Items with negative correlations indicate a scoring key error or serious problems with the item functioning — for example, multiple correct answers, content that was not taught, confusing content, or wording.

4. Distractor-total correlation (selected-response items) – This statistic describes the relationship between selecting an incorrect response for a specific item and performance on the entire test. Typically, the correlation between an incorrect answer and total test performance is weak or negative. The values of this correlation are typically compared and contrasted with the discrimination index. Items where item-total correlations for incorrect answers are stronger than for the correct answer will be flagged.

B. Differential item functioning (DIF)

Another set of analyses typically conducted is differential item functioning (DIF). DIF analyses are essential to help verify that items are fair for all groups of students. We will carry out DIF analyses for all major subgroups of interest to the Consortium (as sample sizes permit) using the generalized Mantel-Haenszel statistic (Dorans & Holland, 1993; Mantel & Haenszel, 1959) and the standardization procedure (Dorans & Schmitt, 1991).

DIF statistics indicate whether there are statistically significant differences in item performance for members of gender or other demographic groups who performed similarly on the entire test. We flag items with statistically significant differences in performance for additional scrutiny to rule out possible biased or unfair content that may have been missed in earlier bias and sensitivity reviews.

Using results of the Mantel-Haenszel analyses, items may be classified into one of three categories based on threshold levels: Category A contains items with no or negligible DIF, Category B contains items with moderate DIF, and Category C contains items with strong DIF. In this system, items classified in Category C are typically submitted for review to assessment.
specialists and/or content data review committees, who consider any identifiable characteristics that may have contributed to the strong DIF.

Other DIF classification systems can be adopted at the discretion of the Consortium. We will work with the Consortium to develop the classification scheme. We will provide items flagged for DIF to the Consortium and provide associated technical documentation.

2. IRT Model Selection

IRT models are ideally suited to the assessments and measurement goals of the Consortium. In particular, mixed item-format tests consisting of the combined scaling of dichotomous (selected-response), short answer, and performance tasks (Yen & Ferrara, 1997; Fitzpatrick et al., 2006) can be accommodated. As demonstrated by years of successful application in K–12 testing programs, IRT models have the flexibility and strength to support the Consortium goals. These IRT models also support adaptive testing (Wainer, 2000).

Within the family of IRT models there are two major choices to be made: 1) use of a unidimensional or multidimensional model, and 2) use of a Rasch one-parameter/partial credit model (Rasch/PC) combination, a two-parameter logistic/generalized partial credit model (2PL/GPC) combination, or a three-parameter logistic/generalized partial credit (3PL/GPC) combination. In this section we will first describe the unidimensional and multidimensional models and the rationale for our recommendation to use a unidimensional model. Then we describe the various unidimensional models and provide evaluation criteria to assist the TAC in recommending the best model and the Consortium in making the final decision.

Unidimensional versus multidimensional models. Operational implementations of IRT models in K–12 testing historically have been unidimensional models. We know that student achievement is not affected by only one trait or ability, and there is no pretense that unidimensional models are identifying or measuring only one trait. In these models, the trait or scale on which items and students are jointly ordered is identified, in essence, as the major dimension that best explains student performance. For Rasch unidimensional models the trait reflects the simple, unweighted combination of the item scores; for the 2PL/GPC and 3PL/GPC models, the trait reflects a weighted average of the item scores, where the weights are functions of the item discrimination parameters. The Rasch/PC and 3PL/GPC unidimensional models have been extensively demonstrated in K–12 testing programs to provide strong horizontal equating (i.e., between test forms within a grade or course) as long as the content and statistical specifications are well maintained for every test form. That is, even though the assessments are not unidimensional, a unidimensional model can provide accurate equating as long as the multidimensionality reflected in the content specifications is essentially consistent across test forms (Reckase, Ackerman, & Carlson, 1988).
Multidimensional IRT models (MIRT) propose that multiple abilities or dimensions contribute to the examinee’s responses to items, and these models estimate additional item parameters and traits (Reckase, 2009). Many different parameterizations of multidimensional models have been formulated and applied for a variety of purposes, such as examining the underlying structure of data, checking model assumptions, or defining diagnostic profiles of student learning.

Multidimensional versions of the normal ogive, Rasch, and three-parameter models have been proposed for multiple-choice and constructed-response items (Ackerman, 1994; Adams, Wilson, &Wang, 1997; Reckase, 1985; Yao & Schwarz, 2006). MIRT models can be characterized as either being compensatory or non-compensatory (Ackerman, 1989; Bolt & Lall, 2003). Compensatory versions allow the dimensions to interact, with high ability on one dimension potentially compensating for lower ability on a second dimension. Multidimensional versions of unidimensional methods have been developed for purposes such as equating and adaptive testing.

In an ideal design where assessments are long, sufficient numbers of items can be administered to provide reliable and valid measurement of multiple dimensions. Unfortunately, without extensive testing, it may be problematic to measure multiple traits with sufficient accuracy necessary for the Consortium. The Consortium design is complex and uses mixed item-format tests, vertical scaling, and computer-adaptive testing, with strong requirements for year-to-year equating. Within this measurement setting, we strongly recommend the use of unidimensional models for scaling those components associated with critical decisions concerning students, teachers, and schools.

Multidimensional approaches offer richer profiles of students’ knowledge states. To this end, we recommend that the Consortium consider conducting several studies to better understand the models and their potential use, for example, exploring the use of multidimensional modeling to report diagnostic subscores in the formative components. An advantage of this approach is that subscores having sufficient numbers of items that can be linked over different test administrations, thereby permitting inferences concerning change.

**IRT Model Options** Both the Rasch/Partial Credit models (Rasch/PC) and three-parameter logistic/generalized partial credit models (3PL/GPC) have been successfully used for decades in many operational K–12 testing programs involving mixed item types. The Rasch/PC model has the advantages of simplicity and a lower case count requirement for item parameter estimation. The 3PL/GPC model is more complex and requires greater sample size, but it has the advantage of greater flexibility. In particular, the item discriminations of dichotomous or multi-level items can be systematically different, and while the 3PL/GPC model can estimate these differences, the Rasch/PC model is restricted to having equal discriminations for the two item types, which can lead to substantial item misfit (Fitzpatrick et al., 1996). This can be a particular issue
when a performance item score is a testlet or the sum of ratings from multiple raters (Sykes & Yen, 2000).

In some instances, the “guessing” parameter (c parameter) of the 3PL model can be difficult to estimate. If that proves to be a serious issue, then either the c parameter can be held fixed at a credible non-zero value (such as 1/A -.05, where A is the number of answer choices) or the c parameter can be held fixed at 0. When the c parameter is held fixed, it reduces to the two-parameter logistic model (2PL). The 2PL can be used in conjunction with the generalized partial credit model in analyzing items from mixed-format tests. While the 2PL/GPC model has not been widely implemented in K–12 testing programs, we anticipate that it would work sufficiently well.

ETS has extensive K–12 experience with both the 3PL/GPC model and the Rasch/PC models. We also have extensive experience with the 2PL/GPC model with international testing programs. For the Consortium assessments with their varied, innovative item types, prior to the completion of the anticipated study we would recommend the 3PL/GPC or 2PL/GPC models, because they place fewer restrictions on the characteristics of items, allowing the content — rather than a statistical model — to more strongly influence item selection. However, we also recognize that the Rasch/PC model is currently being successfully used in many settings, and following the analysis of the data this model may produce similar results and be selected by the Consortium. Regardless of the model selected we have experience using all of the models and will provide all of the analyses that will assist the Consortium in choosing the model best suited to its needs, assessments, and students.

To support the Consortium in the IRT model selection process, using the data collected in the pilot study, we prepare the following information for the Consortium’s TAC to consider:

1. Complete the dimensionality study designed to inform the vertical scaling design as pertains to the degree to which responses support a multiple-construct interpretation. The dimensionality study and associated methodology was discussed in Section 2.4. The results of the dimensionality investigation will be reviewed first and the implications for IRT scaling discussed.

2. Compare the set of potential parameterizations, such as Rasch/PC, 2PL/GPC, and 3PL/GPC based on the vertical scaling data.

3. Examine the residuals, goodness-of-fit analyses, and standard errors that will be used to evaluate the quality of the resulting item parameter estimates. Item non-convergence and other diagnostics will also be examined.

4. Link pilot forms across grades using separate calibration and compare the ordinality of the resulting vertical scales in the context of comparing several candidate IRT models.
5. Finally, the implications for the measurement of growth will be evaluated. As suggested by research, the choice of the IRT scaling model will have the largest impact on measurement of growth followed by the choice of the calibration design and IRT score type (Briggs & Weeks 2009).

Following the model selection the items will be calibrated using either an open-source or commercially available software. We have identified several options for consideration (see Table A.1). The item statistics will then be used for the pilot data review meeting and loaded to the Consortium item bank. The following identifies the item parameters and the associated model or models.

1. All potential IRT models contain b-parameter estimates. These measures of item difficulty are very similar in interpretation to classical item difficulty or \( p \)-value (though scaled differently).
2. Polytomous items will have parameter estimates associated with thresholds between response categories. An item with score categories 1,2,3,4,5 will have four \((k-1)\) thresholds.
3. \( a \)-parameter estimates will provide another measure of item discrimination should the 2-parameter logistic, 3-parameter logistic, or generalized partial credit models be used.
4. If the 3-parameter logistic model is selected, the \( c \)-parameter describes the item lower asymptote in a selected response item.

### 3. Test Form Information

In addition to the item level information we will conduct the following analyses to provide information at the test level:

a. **Classical Test Theory Reliability Estimates.** For fixed-form administrations, where the goal is to estimate the precision of a set of test scores, a measure of internal consistency is frequently used. Measures of internal consistency, such as Cronbach’s coefficient alpha, \( \alpha \) (Cronbach, 1951) and the Feldt and Brennan (1989) statistic may be used for estimating the reliability of the test scores. We will report standard errors of measurement, estimated from the standard deviation of test scores and the test reliability coefficient.

b. **Classification Consistency and Accuracy.** Analyzing and reporting classification consistency and classification accuracy with respect to the performance level cut scores is an important source of evidence. More specifically, for each administration we will produce the following estimates of classification accuracy and consistency (Livingston & Lewis, 1995):

   » **Decision accuracy** estimates the extent to which examinees are classified in the same way as they would be on the basis of the average of all possible forms of a test. Decision accuracy addresses the following question: How does the actual
classification of examinees, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores were somehow known?

» **Decision consistency** estimates the extent to which examinees are classified in the same way they would be based on a single form of a test other than the one for which data are available. Decision consistency addresses the following question: What is the agreement between the classifications based on two non-overlapping, equally difficult forms of the test?

c. **IRT reliability information.** For each pilot test form, we will provide plots of the test information functions (See Figure 4). These graphics will support comparative analyses across administered pilot forms, and may be used to identify target information curves for future administration. In addition, IRT information may be used to evaluate gaps in the item pool relative to the test blueprint at the level of total test, subscore, or content standard.

We are providing examples of such comparisons below at the level of content cluster. ETS will be responsive to Consortium requests for adjustments to these procedures.
**Figure 4.** Plots of test information functions (TIFs) provide a way to easily compare overall and subscore information for each test against target TIFs and other pilot forms.

---

**4. Evaluation and Monitoring of Performance Task Scoring**

There are a number of procedures for evaluation and monitoring of performance assessment tasks, ranging from ones that occur early in the scoring, such as range-finding, to ones that might occur over years that entail rater drift such as trend scoring. In concert with the Consortium and your scoring vendor, we propose to support the scoring process of performance tasks as described below by validating the range finding, monitoring the real time rater agreement, conducting rater trend studies and describe the linking design used to demonstrate task consistency. These studies will provide a basis for monitoring whether the score categories and the underlying construct is maintained into the field test and subsequent administrations.
Validate Range Finding. We assume that the proposed rubrics and range finding responses will be identified by the scoring Contractor, and recommend that the proposed rubrics will be applied by at least two judges to at least 200-300 responses per subject and grade level. The actual number that is required will depend on the form of the performance tasks, characteristics of the automated scoring model, and characteristics of examinee responses to the tasks. We will ask that these judges rate responses to all performance tasks within the form being evaluated, and will request that the scoring Contractor provide the resulting data to us.

During the validation of the range finding, we assume that the content expert panel will review the scoring rubrics, raters’ abilities to consistently apply these criteria, and the ability of the automated scoring model to match raters scoring. We assume that calibration/benchmark papers for any tasks that require ongoing rater scoring will also be identified during these meetings. Based upon the values of these agreement indices, we will work with the Consortium to establish flags to identify items that are not performing consistently with expectations and that require careful review in evaluating whether automated scoring is appropriate for the item/task. We will provide the full complement of rater agreement statistics for use in a meeting with the scoring Contractor and the external content experts to finalize the performance task rubrics, and to identify benchmark student performances.

Some validation of scoring is necessary for automated algorithms. Consistent with the procedures used with rater protocols for monitoring reliability, we can produce the following statistics:

1. Similarity of human and automated score frequency distributions and moments (means and standard deviations).
2. Standardized differences (effect sizes) between human and automated score means. This is computed as the difference between means divided by the standard deviation of the human scores.
3. Unweighted Kappa statistics (Cohen, 1960), which characterize the degree of agreement or association between automated and human scores after correcting for chance agreement rates.
4. Quadratic-weighted Kappa statistics, which have similar properties to unweighted Kappa, but are degraded disproportionately by the presence of large disagreements between some human and automated scores.
5. Pearson correlations, which provide another measure of the degree of agreement or association between automated and human scores.

a. Monitor Real-time Rater Agreement

For those performance tasks that require rater scoring, we propose to address consistency of human ratings during the scoring window. Early in each scoring session, we will ask the scoring Contractor to have each rater score 8 to 10 monitor/calibration/benchmark papers. Several
indices will assess the extent to which a rater is well-calibrated to score. These include the percent of exact agreement, the sum of differences between each rater’s scores and the consensus scores, and the sum of squared differences. All of these indices will be evaluated to determine if remedial action should be taken.

We recommend that following the successful completion of a certification test, each qualified rater receives rigorous training in correctly applying the rubric at each specific score point. These should be documented in the Technical Report. Both during training and throughout the entire scoring process, raters should be guided by the benchmark and range-finder responses selected during the range-finding process, which serve as models of the level of quality required for each score point. We also recommend that after training is complete, raters take daily calibration tests to further establish that they are scoring reliably and consistently; they should be monitored during scoring by scoring leaders who constantly review and analyze raters’ scores on interspersed validity responses that provide ongoing measures of rater accuracy. If the scoring Contractor has this capability, we will regularly analyze real-time data, such as those collected electronically by online scoring systems, to further establish the reliability of the scoring process.

3. b. Assure Categorical Comparability Across Prompts and Across Years: Rater Trend Studies

There are three possible ways that performance tasks might be scored. The approach we propose to evaluate that scoring is different for each scenario.

a) If an automated scoring model is applied to all performance tasks in a particular family, we recommend that the scoring vendor enlist human raters to score a small number of responses at each administration to confirm that the automatic scoring is functioning as expected. The absolute number of such ratings will depend on the form of the performance tasks, characteristics of the automated scoring model, and characteristics of examinee responses to the tasks.

If a difference is observed between the automated and human scoring, we recommend that the scoring Contractor have the same group of raters assign scores to responses from a historical administration. Consistency with historical human ratings will suggest that new human ratings have not changed, and the automated scoring would be suspect for the particular prompt at issue. Inconsistency of new form human ratings with historical human ratings might suggest a training problem, and would therefore suggest that the raters be recertified and the papers be scored again.

b) If the automated scoring approach implemented requires that a new scoring model be trained for each new task, then human scored responses will already have been collected in order to support the training and evaluation of those models for each item. This human-scored response data is sufficient to establish the reliability of scoring, but does not obviate the need for human scoring of a sample of responses in each
operational administration where the item is used, so that discrepancies in item response characteristics can be tracked. The procedure used for resolving inconsistencies between automated and human scoring in the previous case would also apply here.

c) If human scoring of a task or tasks occurs, we would recommend a more in-depth evaluation called a rater trend study. Tate (1999) articulated a solution to the problem of subjective or changing scoring standards for constructed-response items (similar to the procedures followed in Fitzpatrick, Ercikan, Yen, & Ferrara, 1998). He suggested an equating study in which any across-year changes in rater severity could be isolated, so that across-group ability differences could be adjusted, if necessary. The equating study envisioned by Tate involves rescoring responses to the constructed-response items obtained from a previous administration. These responses, obtained from the previous group of examinees, would be scored by the same raters scoring responses for the same items for the new group of examinees. These trend papers have two sets of scores associated with them: one from the old set of raters and one from the new raters which are then compared.

If the Consortium adopts this approach, we will designate one testing cycle as the baseline cycle. Responses from this baseline cycle will be randomly selected and scored during each future cycle. The scores from the future cycle will be compared to the baseline scores to obtain a conversion table of equated scores. Any of several criteria could be used to determine whether scoring standards have changed sufficiently from the baseline year so that it is necessary to use the converted scores. We recommend the use of the conditional standard error of equating (CSEE) from the mean-standard deviation linear equating approach. An asymptotic approximation to the CSEE is given by,

\[
CSEE \approx \sqrt{\frac{S_Y^2(1 - r_{XY})}{N} \left[ 2 + (1 + r_{XY}) \left( \frac{X_i - \bar{X}}{s_X} \right)^2 \right]}
\]

where \( Y \) represents the set of baseline scores; \( X \) represents the new set of scores; \( \bar{X} \) and \( s \) are the mean and standard deviation, respectively, of the scores; \( r \) is the correlation between the two sets of scores, and \( X_i \) is one of the possible score levels for the assessment. If the original score and the converted score differ from each other by more than twice the CSEE at any point, this will be taken as evidence that a non-negligible shift in scoring standards has taken place.

A second optional step consists of new set of ratings being directly equated to the old set using the mean-standard deviation linear equating approach. The outcome is a table listing, for each score from the new raters, the corresponding score if the previous raters had scored the entries (i.e., the equated scores). All scores for the current year are converted to scores from the previous year using this table.
Even if trend scoring is not used, we will summarize the responses for each performance task after scoring is complete. These summaries will include statistics such as:

- Frequency distribution of raw task scores,
- Frequency distribution of equated multiple choice scores, and
- Correlation between task scores and multiple choice scores.

**d. Generalizability Analyses**

The Consortium may also be interested in generalizability analyses of performance task scores to quantify the proportion of variance explained by various possible sources of variation, including raters, task, and persons (desired variance).

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides a useful analytical framework for examining the simultaneous impact of multiple sources of error and their interactions. Cronbach et al. (1972) combined and expanded upon classical test theory and ANOVA techniques in their development of generalizability theory (G-theory). A facet in G-theory is defined as a set of similar conditions of measurement, such as a set of examinees, referred to as persons for this discussion, items or raters (see Brennan, 2001, for an extensive treatment of generalizability theory). The application of generalizability theory is relevant to many problems pertaining to constructed-response items, such as the Consortium’s performance tasks.

Prior to the introduction of G-theory, an important question pertained to whether more variance was attributable to raters or to items/tasks. In a typical G-theory study, different facets of measurement can be identified such as persons \((p)\), raters \((r)\), items \((i)\), and temporal variation or occasions \((o)\). These measurement objects or facets can also be crossed with interaction (persons by items, denoted as \(p \times i\), for instance). Like ANOVA, G-theory allows for fixed and random effects for the measurement facets. G-theory differs from ANOVA in its emphasis on the estimation of variance components, rather than the significance of factors tested using F-tests, in analyzing the contributions of the various facets to overall measurement error. The univariate generalizability model further partitions the single error term of classical reliability theory into separate variance components. The universe score of G-theory is analogous to the true score of classical test theory. G-theory distinguishes between relative and absolute interpretations. For relative decisions all variance components that influence the relative standing of individuals contribute to measurement error. For absolute decisions, all variance components contribute to error except the object of measurement. In a G-theory study, it is desirable to implement a fully crossed design that permits the estimation as many sources of measurement variability as possible.

In large-scale assessments, it is usually not feasible or even desirable to have raters score every examinee (person). In that case, the analyses may include objects of measurement which are nested such as raters within items denoted as \(r : i\).
For example, if two raters scored each student’s response but each student did not receive the same prompt, a nested unbalanced design may be used (Lee & Kantor, 2005; Wang, Zhang, & Li, 2007) as described below:

\[ \text{Design} = (\text{Person} : \text{Task}) \times \text{Rater} \]

The model assumes that the raters are selected from an infinite pool of raters and all the raters are randomly equivalent. The model also assumes that the writing prompts are randomly selected from a universe of prompts and that students’ writing responses are randomly assigned to the raters.

We look forward to working with the Consortium and its technical advisors on the design of a set of studies to meet the goals for consistency in performance task scores.

e. Linking design to demonstrate task consistency

After rater-monitoring procedures have occurred to ensure that tasks are scored reliably and meet the established criteria, they will be placed on an IRT vertical scale using the proposed linking design described in Section 2.1.1. This approach facilitates comparison of performance tasks within and across administrations.

Should it not be possible to scale the performance tasks, the equivalent groups aspect of the data collection design described in Section 2.1.1 will provide a means to evaluate consistency of performance task scores for the pilot administration. Those responding to each performance task are sampled to be randomly equivalent, so that differences in performance will reflect task differences once rater effects are taken into account (see Section 2.7, part 4.c. above for a description of an adjustment for rater harshness). For future use, any differences in difficulty across performance tasks can be adjusted.

The evaluation of performance task consistency will also include measures such as frequency distributions of task scores and correlations between task scores and equated multiple choice scores.

5. Accessibility for ELs and SWDs

Recent changes to assessment design (universal design, evidence-centered design, access by design, and online delivery, with accessibility features such as read aloud) have most likely contributed to improved accessibility of state assessments for English learners (ELs) and students with disabilities (SWD). The degree to which these changes were made varies across states and, to some extent, disability subgroups. For example, the addition of an audio button on Mathematics assessments would most likely improve the accessibility (and performance) of students with reading-based learning disabilities, but is less likely to have an impact on students with visual impairments if other accessibility features (magnification, refreshable Braille, voiced keyboard navigation) are not included in the online platform. These interactions between the
accessibility features/accommodations and student characteristics make it necessary to use research designs to provide evidence that:

- Accommodations and accessibility features used on Consortium assessments are effective in making assessments more accessible for ELs and SWDs, while also
- The resulting test scores are a valid measure of the intended construct.

We propose a framework to answer two main questions:

- Does the accommodated test score accurately measure the construct being assessed for students with disabilities than students without disabilities?
- Are the accommodated test scores for students with disabilities psychometrically comparable to the standard test scores for students without disabilities?

Large experimentally designed studies (e.g., differential boost or interaction hypothesis) are appropriate for these questions. However, there are limitations of cost and feasibility (e.g., when students cannot test under non-accommodated conditions) so, several other research designs are possible. These include field testing new item types and test forms on a sample of students with disabilities, and cognitive labs or think aloud studies. In addition, we will examine the comparability of test scores using pilot test data and an accommodations sample (oversampled to include a large number of students with a variety of disabilities, English language proficiency levels, and native languages). Types of analyses to be included, when sample sizes allow, are: (1) test reliability, (2) factor analysis, (3) differential item functioning, and (4) predictive validity. These forms will be consistent with the paper administration.

We will work with the Consortium and its Technical Advisory Committee to develop a research agenda to examine the effectiveness of the accommodations and accessibility features used in the Consortium assessments, and the degree to which they support fidelity to the focal constructs and provide valid assessment outcomes.

We are currently engaged in a number of projects focused on technology-based assessments for students with disabilities (e.g. testing on user-owned devices and assistive technology, including refreshable Braille capabilities for testing; a new method of reading aloud Mathematic; and a study to examine testing on an iPad® and netbooks for students with and without visual impairments). Research on technology-enhanced assessments for students with visual and/or fine motor impairments is also being conducted.
Finally, we have collaborated on research on accessibility features and testing accommodations for ELs and SWDs, including:

- Psychometric properties of K–12 assessment, such as factor structure, item functioning, and reliability (Cook, Eignor, Sawaki, Steinberg, & Cline, 2010; Cook, Eignor, Steinberg, Sawaki, & Cline, 2009; Laitusis, Maneckshana, & Monfils, 2009; Steinberg, Cline, Ling, Cook, & Tognatta, 2009; Stone, Cook, Laitusis, & Cline, 2010)
- Impact of testing accommodations on growth models (Buzick & Laitusis, 2010; Buzick & Laitusis, 2011)
- Large print and Braille accommodations (Stone, Cook, Cline, & Cahalan-Laitusis, 2010; Bennett, Rock, & Novatkaski, 1989)
- Audio presentation of test content (Hansen, Shute, & Landau, 2010)
- Calculator use (Bridgeman, Harvey, & Braswell, 1995; Bridgeman, Cline, & Levin, 2008)
- Extended time (Cahalan-Laitusis, Mandinach, & Camara, 2002; Cahalan-Laitusis, King, Cline, & Bridgeman, 2006; Mandinach, Bridgemen, Cahalan, & Trapani, 2005; Bridgeman, Cline, & Hessinger, 2003)
- Screen size (Bridgeman, Lennon, & Jackethal, 2003)
- Comparability of computer and paper-based tests (Gallagher, Bennett, and Cahalan, 2001; Schaeffer et. al., 1998)
- Word processed vs. handwritten essays (Powers, Fowles, Farnum, & Ramsey, 1992; Powers, 1999)

6. Comparability study of online and paper/pencil versions

The final study is the comparability study of the online and paper and pencil versions. The data collection design described in Section 2.1 was developed to provide a sample that could be used for this study. It is not uncommon as a testing program moves to an online modality that the tests are also administered in paper and pencil where there is limited access to electronic platforms. The Consortium assessment, by design, may increase the potential mode effect. These include:

- Technology enhanced items or items scored automatically that are not amendable to paper versions;
- Ability to return to a question and change an answer is unavailable on some online but available on paper; and
- Maximum time for testing is typically static for paper, but is effectively variable in some other forms of online administrations.
While it will not be possible to mitigate all of the differences between items administered online and on paper and pencil versions, we do recommend that this issue be considered during the form construction process. Fortunately, we can mitigate the impact resulting from some of the other potential sources of differences. An important component of minimizing the differences between the two different modes is through the administration procedures. Therefore, we would propose to work closely with the Consortium and your appropriate Contractors to review the administration procedures and make recommendations, if necessary. Some of these best practices we would recommend include:

» Including a very short tutorial in the online test, to avoid computer unfamiliarity/inexperience adding an irrelevant source of difficulty for the online version;

» Establishing consistency in presentation format of long passages between online and paper, to support a similar reading experience;

» Establishing testing conditions are as similar as possible. This may include holding paper administrations in computer labs and maintaining a similar level of test security, testing time; and

» Establishing that a clock is similarly available to online and paper administrations, to control for examinees’ time management.

We will also conduct a mode comparability study. For this study, the goal is to select samples taking each version that are as equivalent as possible. Paper forms will be administered to selected, representative samples of approximately 1500 students each. As mentioned previously in Section 2.1, a subset of schools could identify two classrooms per grade to be tested — one for online administration and one for paper administration; it would be ideal if students from these classrooms could then be randomly assigned to testing mode. A fallback approach is to identify matching schools and assign one from each pair to a mode. Propensity scoring could also be used to match schools for the two samples. A computer version corresponding of the paper form will be among the forms spiraled to the on-grade level online examinees that will facilitate the mode comparability analyses. A content representative test form, which includes both summative items and performance tasks, will be selected for use in the study.

A computer familiarity questionnaire will also be administered to both groups; this will include questions about computer access, use, and experience, as well as test preparation. We will make initial comparisons of score distributions of paper and online modes using descriptive statistics. Because identical forms will be used across delivery mode, it will be possible for us to use differential item functioning (DIF) analyses to evaluate delivery mode comparability at the item level. DIF analyses are typically used to help verify that items perform in the same manner for demographic groups. We propose to carry out DIF analyses of selected response and constructed response items for online versus paper administrations using the Mantel-Haenszel (MH) statistic (Dorans & Holland, 1993; Mantel & Haenszel, 1956) and the standardization
procedure (Dorans & Schmitt, 1991). We will analyze data to evaluate DIF for delivery mode concerns.

For this study, we will separately calibrate and link paper forms to the online forms, using all online items as the linking (anchor) set. This is very consistent with method used for the annual equating of tests, except for anchor length. Analysis of comparability at the item level will include comparison of item parameter estimates using IRT-based anchor item outlier detection methods (item stability). In the outlier analysis, this process involves an inspection of differences between the paper parameters estimates and the online ones for the linking items. Items with large weighted root-mean-square differences (WRMSD) between item characteristic curves (ICCs) based on the online and paper difficulty estimates are eliminated and the linking constants are re-estimated. The weighted differences are calculated as follows:

$$WRMSD = \sqrt{\sum_{j=1}^{61} w_j \left[ P_n(\theta_j) - P_r(\theta_j) \right]^2}$$

where $\theta_j$ ranges from -3.0 to 3.0 by 0.1, $w_j$ is a weight equal to the proportion of estimated abilities from the transformed paper form in interval $j$, $P_n(\theta_j)$ is the probability of correct response for the transformed new form item at ability level $j$ and $P_r(\theta_j)$ is the probability of correct response for the “reference” form (i.e., the online estimates). Any items in which the difference WRMSD is greater than 0.125 are eliminated from the linking set.

The differences observed between modes should be comparable to those observed between horizontally equated tests. Test level analyses will include an examination of psychometric properties for tests delivered in paper and online modes such as the comparison of scoring tables.

The detailed steps for the comparability study would consist of the following:

1. Select forms that assess the entire test blueprint for online and paper groups;
2. Obtain matched student samples for online and papers forms;
3. Calibrate the online test forms; calibrate the paper test forms;
4. Use all online items as anchor and place paper form onto the “online” scale;
5. Perform anchor stability analysis (e.g., WRMSD); items functioning very differently across mode will be dropped from the anchor linking set;
6. Apply anchor linking method using the final linking set from step 5 above (e.g., Stocking & Lord, 1983);
7. Plot the resulting item parameters by mode and evaluate departure from the identity line;
8. Compare resulting scoring tables (raw score to scale score) by mode; and
9. Compare scale score distributions across mode and evaluate differences.
We will report these investigations to the Consortium, which will identify risks and recommend solutions, should they be needed. If the Consortium and the TAC wish, we would be pleased to discuss options for the proposed analyses or additional analyses.

**Proposed Format for Committee Review**

We propose that the pilot form be submitted as a set of electronic item cards rather than as a composed test form. Evaluating test content at the item rather than the form level will become increasingly necessary once adaptive testing begins.

We recommend that item cards include the item itself and any information related to the item’s content on one side of the card. We recommend that item statistics (classical and IRT) and differential item functioning (DIF) results appear on the other side of the card. Information from the online/paper comparability study will appear as an additional DIF comparison.

We will include other information as well, if desired. These formats are easily configurable.

**Proposed Format for Item Bank Database**

We will begin with an approach that resembles the one described in Section 2.6 when developing and delivering upload files containing item statistics. We will evaluate the Section 2.7 requirements in conjunction with Section 2.6, and determine if there is commonality in the data deliverable and whether or not we can use the record layouts as-is or with minor modifications. In this case, we will still use the file formats proposed in Section 2.6 and reuse them to meet the requirements contained in Section 2.7.
Figure 5: Sample Item Review Card. Item review cards display a typical format for item cards, but the layout is completely configurable to meet the needs of Consortium.

<table>
<thead>
<tr>
<th>ITEM ID: 223232</th>
<th>ETS Item ID: MDA50028</th>
<th>Response Type: SPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaled Item Admin Type: F</td>
<td>Scaled Admin Month/Year: May/2008</td>
<td>Scaled Form &amp; Seq #: 008/21</td>
</tr>
<tr>
<td>Last Item Admin Type: O</td>
<td>Last Admin Month/Year: May/2009</td>
<td>Last Form &amp; Seq #: 089/17</td>
</tr>
<tr>
<td>ETS Standard Code: E3.1.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment Limits: 3.1.3b theoretical probability</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Item: MDA50028

A spinner is divided into 8 equal sections as shown below.

What is the theoretical probability that the arrow will not land on an E or an O?

Proposed Software

A wide array of software is available that might be used to meet the Consortium’s needs. Our staff is proficient in a wide variety of commercially-available and open-source psychometric software programs. Software we believe could be useful to the Consortium is described in Tables A.1 through A.3. We have put these software packages into rank order according to our evaluation of how effective they are likely to be for Consortium work and by their function. We recommend those with an overall rank of one as being likeliest to support consistent and accurate results. We will work with the Consortium to create an effective solution with any of the software packages listed, or other software that is preferred by the Consortium and its Technical Advisory Committee. We will supply the software code associated with any program as part of the technical documentation, as noted in Section 2.2. We are also willing to work with other software packages that we have not listed if its psychometric properties have been established.
Table A.1. Software for data manipulation, raw scoring, classical item analysis, and differential item functioning.

<table>
<thead>
<tr>
<th>Proposed Software</th>
<th>Software Availability</th>
<th>Comments</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>Commercially available</td>
<td>SAS is an extremely capable and flexible software package. Because the input is a parameter file, we will be able to publish all program code for IA, DIF in the technical report for easy replication at any time. Because SAS is currently used in many organizations, it would be relatively easy for the Consortium to identify staff to replicate our work. We propose SAS for use in performing data manipulation, item scoring, classical item analysis, and DIF.</td>
<td>1</td>
</tr>
<tr>
<td>R (base module)</td>
<td>Open source</td>
<td>R (base module) is well known in research and academic applications, with many software modules making it nearly as flexible as SAS. Because the input is a parameter file, we will be able to publish all program code for item analysis and DIF in the technical report for easy replication at any time. Because R (base module) is currently used in many organizations, it would be relatively easy to identify staff to replicate our analyses.</td>
<td>2</td>
</tr>
<tr>
<td>SPSS</td>
<td>Commercially available</td>
<td>SPSS is similar to SAS in many respects. However, we are aware of no characteristics relevant to this proposal for which it is superior to SAS. It also appears to be less often used in operational settings, which could generate problems as the program is shared from one Contractor to another.</td>
<td>3</td>
</tr>
<tr>
<td>jMetrik</td>
<td>Open source</td>
<td>A specialized item analysis package that appears to perform an array of item analysis and DIF procedures well. More analyses are required to investigate suitability for use in the context of high-stakes large-scale assessments.</td>
<td>4</td>
</tr>
<tr>
<td>Iteman</td>
<td>Commercially available</td>
<td>Iteman is a specialized item analysis package that can handle partial overlap of test forms by creating a criterion score that is based on a subset of items. As with jMetrik, more analyses are required to investigate suitability for use in the context of high-stakes large-scale assessments.</td>
<td>5</td>
</tr>
</tbody>
</table>
Table A.2. Software for IRT equating and scaling.

<table>
<thead>
<tr>
<th>Proposed Software</th>
<th>Software Availability</th>
<th>Comments</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARSCALE</td>
<td>Commercially available</td>
<td>PARSCALE is likely the most widely-used software in operational programs that is capable of calibrating items using all of the commonly-used unidimensional IRT models. We believe that most of the major testing companies use PARSCALE in at least one program, which speaks to both its quality and the flexibility, and it will give the Consortium the ability to easily transition data from one Contractor to the next. PARSCALE has been widely researched, which provides confidence that the results it produces can be relied upon.</td>
<td>1</td>
</tr>
<tr>
<td>IRTPRO</td>
<td>Commercially available</td>
<td>IRTPRO is a relatively new software package that can calibrate items using the IRT models supported by PARSCALE, as well as multidimensional models. IRTPRO can also perform multi-group IRT analyses, which could be helpful, were there to be state-to-state variation in item performance due to variations in curricula. However, IRTPRO is relatively new and has had limited use in operational settings. Our view is that IRTPRO is currently research software that we hope will be ready for operational use soon. However, at this time, we believe more analyses are required to investigate suitability for use in the context of high-stakes, large-scale assessments.</td>
<td>2</td>
</tr>
<tr>
<td>WINSTEPS</td>
<td>Commercially available</td>
<td>WINSTEPS is a well-known and widely-used software package for implementing Rasch models. Other models outside the Rasch family cannot be evaluated using it.</td>
<td>3</td>
</tr>
<tr>
<td>ICL</td>
<td>Open source</td>
<td>ICL is software that can calibrate items using most of the IRT models supported by PARSCALE, including 1-, 2-, and 3-parameter logistic models, partial credit, and generalized partial credit. ICL can also perform multi-group IRT analyses, which could be helpful, were there to be state-to-state variation in item performance due to variations in curricula. Our view is that ICL is currently research software and more analyses are required to investigate suitability for use in the context of high-stakes large scale assessments.</td>
<td>4</td>
</tr>
<tr>
<td>Proposed Software</td>
<td>Software Availability</td>
<td>Comments</td>
<td>Rank</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------</td>
<td>----------</td>
<td>------</td>
</tr>
<tr>
<td>R (ltm)</td>
<td>Open source</td>
<td>R (ltm) is software that can calibrate items using most of the IRT models supported by PARSCALE, including 1-, 2- and 3-parameter logistic models, as well as the generalized partial credit model. Our view is that R (module ltm) is currently research software that lacks sufficient evidence to support its operational use. At this time more analyses are required to investigate suitability for use in the context of high-stakes, large-scale assessments.</td>
<td>5</td>
</tr>
</tbody>
</table>

**Table A.3.** Additional psychometric software.

<table>
<thead>
<tr>
<th>Function</th>
<th>Proposed Software</th>
<th>Software Availability</th>
<th>Comments</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRT Equating</td>
<td>STUIRT</td>
<td>Open source</td>
<td>STUIRT is software that conducts IRT scale transformations for mixed dichotomously and polytomously scored tests. It supports all of the unidimensional IRT models supported by PARSCALE. This software was developed at the University of Iowa, and has been used in statewide testing programs, including Maine, Maryland, Michigan, and Oklahoma.</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>IRTEQ</td>
<td>Open source</td>
<td>IRTEQ is software that conducts IRT scale transformations for mixed dichotomously and polytomously scored tests. This software appears to be as capable as STUIRT, but is not preferred due to the record of successful use that STUIRT has.</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Plink</td>
<td>Open source</td>
<td>Plink is software that conducts IRT scale transformations for mixed dichotomously and polytomously scored items. Its properties are not well-known.</td>
<td>3</td>
</tr>
<tr>
<td>Generalizability</td>
<td>GENOVA</td>
<td>Open source</td>
<td>GENOVA is widely-used, free software associated with Robert Brennan, the scholar most closely associated with generalizability theory. The program has been applied to the conduct of generalizability and dependability studies for over 20 years.</td>
<td>1</td>
</tr>
<tr>
<td>Function</td>
<td>Proposed Software</td>
<td>Software Availability</td>
<td>Comments</td>
<td>Rank</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------</td>
<td>------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Dimensionality</td>
<td>BMIRT</td>
<td>Open source</td>
<td>BMIRT is multidimensional IRT software package that functions in both in a confirmatory mode (items are specified to be associated with factors) and in an exploratory mode (items association with factors is determined through a statistical evaluation). This makes BMIRT potentially useful in evaluating dimensionality for the Consortium.</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>LISREL</td>
<td>Commercially available</td>
<td>LISREL is a well-known and widely used structural equation modeling software package capable of performing dimensionality evaluations with complex data. This includes the dichotomous and polytomous response data that will be associated with the Consortium.</td>
<td>2</td>
</tr>
</tbody>
</table>
2.8 **Conduct Analysis of Pilot Test**

The ETS psychometric team will work with the assessment development and the delivery and scoring Contractor(s) to produce a report summarizing recommended changes for the field test development process, based upon the pilot test data analyses. Upon receiving final committee feedback from the assessment development Contractor, we will produce a report detailing recommendations for changes to the field test development process.

**Item development process revision recommendations**

After administration of the pilot test, we will work cooperatively with the assessment development Contractor and the Consortium to determine if any changes need to be made to the item development process. Some areas we will examine include the implementation of innovative items, the performance of students in each standard, and the distribution of score points for each performance task. The use of innovative items especially will require a careful examination of how students interacted with the technology associated with the delivery system. If our analysis determines that improvements could be made in any area of item development, then we will work closely with the Consortium and the item development Contractor to identify and make any potential improvements in the process. Because a vertical scale will be developed, a plan for item usage and performance across grade levels will also be evaluated.

**Recommendations for revisions to administration materials and processes**

An evaluation of students’ testing environment and their typical computer use can also help inform the field test development process. ETS proposes to create an online survey to gather feedback from examinees about test site conditions and overall testing environment. Analysis of the survey data will be used to inform our recommendations for enhancing and improving administration materials and test site administration processes. Some assessment management systems have an online survey/evaluation tool that allows administrators to ask questions to obtain feedback from examinees on the testing environment. We can also add questions to inquire about the testing process and directions given to students. If the assessment management system allows survey administration, then we propose to submit a survey to be administered through the assessment delivery system. If the delivery system cannot administer a survey to pilot test participants, then we propose using a stand-alone Internet-based survey tool to perform this function. A cross tabulation of results can be produced to evaluate student testing experience and accessibility among the pilot sample and subgroups.

**Analysis of anchor pools available for Field Testing**
As part of the analysis of the pilot test data, we will summarize the psychometric properties of the anchor pools, including but not limited to providing distributions of classical and item response theory (IRT) item and task parameter estimates, fit statistics, and differential item functioning (DIF) statistics. Because a vertical scale will be implemented, the grades in which the item was administered will be included. An indication of the item delivery mode will be included. We will evaluate the extent to which the statistical distributions of items and performance tasks meet the needs of the field test design. We will obtain from the assessment development Contractor their content-balance evaluation of the pilot test anchor pool. We will also base our recommendations for the field test development on traditional measures of reliability and validity used for fixed form tests, as well as test information functions and differential test function information, and the extent to which this statistical information matches the expected outcome for the pilot tests.

**Survey of Pilot Test participants**

A survey of test administrators and students will solicit feedback on the usability, test experience, clarity of instructions, responsiveness of test delivery software, technical ease of use, and other factors affecting test delivery. The Test Delivery Contractor will deliver the survey and collect results. We will work with the Test Delivery Contractor to analyze and report the results. These results will also be used to inform the field test development process.

**Recommend changes to performance task scoring or data collection processes**

Based on the statistical properties of the pilot tests on the performance tasks, and the feedback from experienced scoring experts on the kinds of responses generated by these tasks (please see our response to Section 2.2 for a description of the kinds of analyses we conduct of pilot test results), we will provide formal recommendations regarding any proposed changes to scoring or data collection processes. In addition, we will provide the assessment development Contractor with performance task summary statistics. We typically look at the distribution of item difficulty to help in determining whether the item pool has sufficient depth to cover the needs for future administrations. Looking at the distribution of students across rubric score point values also provides information for evaluating whether the rubrics appropriately delineate the range of student responses. The distribution across score points may indicate that the rubric for that task needs to be rewritten or that the task does not sufficiently assess the skills required for students to demonstrate what they are able to do with respect to the construct being measured.
Another important element with respect to rater scoring of performance tasks will be the level of rater agreement that is present. For large, complex tasks at the upper grade levels, less agreement may be expected. However, items must meet some minimum for rater agreement. The agreement can be tabulated against other raters or accuracy measured against a criterion (e.g., calibration paper). Some performance task types may simply be more problematic to score. In these cases, more resources may need to be marshaled in order to improve rater performance. These resources could include more calibration papers targeted at the problem and/or changes in anchor and training materials. In other cases, where automated scoring (machine scoring) is used, the model (algorithm), electronic rubric, or training set may be in need of modification. Various types of statistics and review will help to identify problems. Our psychometric staff will take all of this information into account when discussing and providing recommendations for changes to the field test development process.

2.9 Design Item and Student Sampling Plans for Field Test and Select Anchor Items and Tasks for Calibration and Building the Vertical Scale

The main purpose of the field test is to calibrate all of the available items on large-scale samples under near-operational conditions in preparation for the operational program, and to validate, adjust, and strengthen the horizontal and vertical scales established based on the pilot test data. We propose to conceptualize the field test as having two phases. In the first phase, large numbers of students will receive linear fixed forms that are reflective of the overall test blueprint consisting of an anchor set of items that remained unchanged from the pilot test, and a field test set composed of new or revised items, along with one or more performance tasks. These data will be used to obtain operational item parameters for the pool of available items, evaluate the performance tasks, and validate, adjust, and strengthen the scales.

In the second phase of field testing, large numbers of students will receive computer adaptive forms that follow operational blueprints, thereby allowing the Consortium to test the online delivery platform for bugs, help assure that the algorithm and item pools are working appropriately, and verify that the platform is functioning as intended. The SBAC program is technically complex and will have several Contractors. To efficiently and effectively execute each phase of development in a manner that best meets the Consortium’s goals and objectives, ETS will work closely with all partners to document each phase of the project to identify needed Contractor hand-offs (e.g., data files, requirements documents, etc.) and key deliverable deadlines to keep the project on track. We will be active partners with the Consortium to anticipate and resolve issues encountered, working with all other Contractors, work groups, the Consortium, and the TAC to help establish that the test delivery at the field test stage adheres to the design for vertical and horizontal linking. We recognize that as the Psychometric Services partner, it will be essential to thoroughly
understand each phase of the program’s development and to coordinate our efforts with the other Contractors.

The specific tasks and our plan for implementation are described below.

1.) Developing the CAT Item Pools.

   a. **Phase 1: Linear Fixed Forms.** In order to create the CAT item pools that will be used for the operational forms, the available item pool for field testing must be administered in linear fixed forms along with an anchor set to calibrate the items and place the items on scale. The steps below describe the linking and sampling design, form assembly (including anchor item and task selection), and the analysis.

   i. **Linking Design.** While the choice of a data collection and linking design is partly contingent on the results of the pilot test and will be determined in conjunction with the Consortium and the TAC, at this time we propose that we use the same hybrid model that is described in Section 2.1.1, incorporating the strength of the common-item and randomly equivalent groups linking designs. That is, assuming that we use a similar linking design, ETS will randomly assign students in each grade level to one of two groups:

   » **On-grade samples:** Students in this group will be administered forms that include items consistent with their grade level. These forms will have some anchor items in common across forms. These anchor items will be used for horizontal equating, placing the on-grade level new and revised field test items on the same scale.

   » **Off-grade samples:** Students in this group will be administered forms that include:

   - items consistent with their grade level (on-grade level items) from the pilot administration, as described in the previous bullet
   - items from an adjacent lower, but not higher, grade (off-grade items), which will be used to validate, adjust, and strengthen the vertical scale.

As with the pilot test, there are two possible configurations for administering the forms. In the first vertical scaling option, all students are administered both a form of items and a performance task component. The advantage of this design option is in the strength of the linking, but it can also be a disadvantage in terms of burden to teachers and students.

In the second vertical scaling option, most students are administered *either* a form or a performance task component. The advantage of this design is that it is intended to reduce the testing time required for most students. The disadvantage is that this design requires more samples and students.
Either of these approaches will provide sufficient data to link the field test forms horizontally and to adjust, validate, and strengthen the vertical scale. In addition, these two designs are flexible and can be amended as necessary in collaboration with the TAC once the results of the pilot test are known and the performance task demands are better understood.

We understand that content and delivery platform concerns may place restrictions on the final configuration of the forms for the administrations, and will advocate on behalf of psychometric considerations (particularly the vertical and horizontal scaling design) to arrive at the best possible solution. ETS will work with the Consortium and its Contractors prior to the start of the field test administration to carefully identify the specifications that will result in the delivery of the horizontal and vertical linking items that will conform to the design requirements for administration and data collection. We will work efficiently and with the foresight that comes from long experience in partnering with organizations of all kinds in the successful achievement of assessment goals. We assume that this process will be done via e-mail, conference calls, and videoconferencing.

ii. **Sampling.** In order to help assure the quality and validity of the item parameters obtained during the field test, we propose to obtain approximately 1,500 students, representative of the target population, by implementing a sampling plan for the first stage of the field test that is similar to that used for the pilot test. A stratified multistage probability sampling method was proposed to select participating schools within each state, with oversampling of students with disabilities (SWDs) and English learners (ELs) for special studies and scaling of special forms. Additional variables of interest defined in collaboration with the Consortium and their TAC would be used in identifying the schools for recruitment. For the adaptive stage of the field test, a similar sampling plan could be implemented to invite participation among those schools with the appropriate technology. However, if many of the schools in the Consortium want to participate in this phase of the field test, we can sample within the resulting participants to achieve a representative sample.

iii. **Field Test Form Construction.** ETS will collaborate with the Consortium and the assessment development Contractor to design field test forms for the first phase of the field test composed of linear fixed forms for the online administration and associated performance tasks. These forms will include items and tasks with known item statistics that remained unchanged following the pilot test — these will form the anchors that will be used to link the forms vertically and horizontally. Each of the forms will also include a set of field test items, which together with the anchor set will reflect the overall test blueprints as much as possible. However, in this stage, preference will be given to helping
assure a solid linking design. Given that the anchor items and tasks will be used to finalize the scales and populate the item pool for the operational CAT, it is critical that the anchor items and tasks associated with the fixed linear forms are of the highest quality possible. We propose to include the following actions, to help assure that the anchor items and tasks are of the best possible quality:

» **Criteria for anchor item selection.** ETS will use our discussion with the assessment development Contractor and the Consortium and its Technical Advisory Committee to determine requirements for the field test administration. To start the development process for the fixed form stage of the field test, we will use the following information to inform the recommendation:

- information concerning the nature of growth and learning sequences demonstrated in the vertical scale established during the pilot
- vertically articulated anchors that must be ordinal in difficulty across grade levels to demonstrate increased expertise
- the number of pilot test items that remained unchanged and are available to function as anchors or linking items
- the content and psychometric requirements for anchor blocks and linking blocks
- items that have some type of dependence (i.e., reading passage dependence) that must be selected together
- A discussion of item acceptance rules that will be needed for the online and subsequent adaptive portion of the field test. These criteria typically include:
  - $p$-values/IRT $b$-parameter estimates
  - item-test correlations/IRT $a$-parameter estimates (if applicable)
  - omit rates
  - DIF
  - item fit

Based on this information, we will make a recommendation for consideration to the Consortium advisors. Our preliminary recommendation is to construct robust anchor sets that comprise 20 percent to 25 percent of each fixed linear, or no fewer than 15 to
20 items. These anchor sets would consist of items that met the minimum criteria for acceptance (as defined above), cover the full range of difficulty, and are as representative of the item types and test content as possible.

2.) Establish composition of anchor item/task selection committees and review committees.

In order to establish the highest possible quality anchor item and task sets, we propose using a small group of individuals to select a recommended pool of anchor items and tasks for each subject area. The two selection committees (English Language Arts/Literacy [ELA] and Mathematics) will include individuals with extensive content knowledge in either the common core area of ELA or that of Mathematics, across the grade levels, and psychometric experts with a deep understanding of the item statistics associated with the pool of eligible content. The committees will provide recommendations, along with annotations about why particular items are recommended, to a larger review committee composed of content experts from the Consortium member states and the ETS psychometricians who participated in the initial selection process. The recruitment and selection of the committee members are described in the next section.

3.) Recruit anchor item and task selection committees.

In collaboration with the Consortium, ETS will identify candidates for the item/task selection committees who have the required deep content knowledge across grades, or appropriate psychometric knowledge, to make strong recommendations for anchor items and tasks. We will prepare accompanying documentation concerning the selections of the committee members and propose candidates for committee membership to be recruited from within the Consortium and other potential stakeholders. After obtaining approval from the Consortium, we will recruit the chosen candidates and provide a description of the group’s item/task selection responsibilities. We anticipate that the group members will complete their work through a combination of individual work and teleconferencing.

4.) Determine decision-making criteria regarding review committee feedback.

ETS will work with the Consortium to develop an efficient methodology for acquiring and evaluating feedback from the larger anchor item and task review committee of stakeholders. We will establish specific written item content and statistical criteria to guide the process of evaluating review committee feedback and arriving at final recommendations based on the criteria finalized above. Suitability for computer administration will be among the criteria considered. Our recommendations will be accompanied with a description of how we arrived at them.

5.) Determine and implement final review/approval process by Consortium.
ETS will submit recommended anchor items and tasks to the Consortium for its review and approval. We will provide the items and tasks with their data, along with the selection criteria, summary of review committee results, and any necessary annotations, to aid the final review and approval process. We will also hold acceptable alternative items in reserve should the Consortium request additional options, changes, or substitutions.

a. **Calibrate new and revised items against the anchor items during the fixed-form phase of field testing.** In this section, we first describe the calibration of new and revised items against the anchors. We then discuss evaluation of the vertical scale, followed by the calibration and linking of the new items to the vertical scale.

**Item calibration against anchor items:** Following receipt of student response data from the fixed form online administration, items and tasks will undergo classical test theory, differential item functioning (DIF), and item response theory (IRT) analyses as described for the pilot in Section 2.7. Item/task statistics will be evaluated against approved criteria, with flagged items sent for review by the Consortium and its test development vendor. All items determined to have satisfactory psychometric properties will then be linked to the vertical scale by horizontal equating to their respective grade scales using the embedded pilot anchor items. Anchor items with large changes in item parameter estimates between pilot and field test administrations will be subject to review and possible removal from the anchor set. The approach used for horizontal linking will be that approved by the Consortium and its TAC.

**Evaluation of the vertical scale:** The basic approach is to cross-validate the vertical established in the pilot. The field test design described above will provide a rich set of items to be used for evaluating the vertical scale and strengthening it as needed. The vertical scale established during the pilot will be evaluated with respect to the ordinality of IRT test characteristic curves across grades, and the stability of between grade relationships. Each grade level will have a TCC based on the IRT parameter estimates obtained. The TCCs should increase monotonically by grade and avoid any disordinal relationships (overlap) between curves, as shown previously in Section 2.1.1. The degree of separation between TCCs will partly be a function of the growth demonstrated in the vertical scale. Curves for the conditional standard error of measurement (CSEM) also need to be evaluated. In general, these curves should provide measurement over a range of ability where
error is minimized. The CSEM should be low and flat and also
display ordinal characteristics. In this case due to the properties of
vertical scales, the CSEM curves will overlap to some extent, which
is acceptable. If these properties and relationships fail to hold for
the scale established during the pilot, then the vertical scaling can
be repeated with the cross-grade linking items administered during
during the field test.

b. Phase 2. Monitor the functioning of the adaptive testing algorithm during the
adaptive phase of field testing.

Once a sufficient number of fixed linear forms have been administered, and the
resulting calibrated items added to the pool, tests using the computer adaptive
algorithm can be deployed. Ideally, prior to the time that students begin
receiving computer adaptive forms, we will be able to use simulation studies to
test the adaptive testing algorithm and the item pools intended for field testing.
As computer adaptive testing events go live in the field test, we will monitor
how they are functioning on an ongoing basis. We will need to work with the
adaptive test delivery Contractor to develop a system for obtaining all
information needed to evaluate the adaptive testing system at various points
throughout the field test (e.g., criteria, specifications, or constraints that the
algorithm is attempting to meet; student item response strings; and resulting
ability estimates produced, as well as the composition of each item pool used
during field testing). We describe the details of our plans for verifying the field
test computer adaptive forms in Section 2.10, and include specifics on how we
will determine how each simulated and real test event meets the test blueprint
and specifications. In Section 2.10, we also outline how we will analyze the
simulated data to report measurement errors for overall scores and subscores,
and how the adaptive testing algorithm using operational item pools can
recover simulated examinee ability, including analysis of estimation bias and
random error. ETS will also analyze the adequacy of each item pool by
examining overall and subscore content, item statistics, plots of test information
functions, and plots of differences in information from the “average” item pool
in that content area and grade level. We will also analyze the results of student
performance on the adaptive tests including those with accessibility features. In
particular, we will include score distributions overall and by subgroups of
interest to the Consortium. Please refer to Section 2.10 for a full description of
how we will monitor the adaptive testing algorithm.

Following the completion of the fixed-form and CAT stages of the field test
administration, all items will be recalibrated and will be scaled onto the final
operational scale using only field test responses. As specified in the Q&A, the
Consortium’s expectation is that items calibrated during the pilot phase will
retain their parameters. Therefore, parameters will only be updated if the pilot item parameters appear to have shifted after field test administration. We will provide a report to the Consortium on items that were recalibrated and parameters updated due to shift.

6.) Performance Tasks

In other sections of this proposal, such as Section 2.7, we describe the types of item statistics that we propose to use to evaluate task properties and rater consistency. Some of the standards that we described earlier for empirical evaluation are listed below. While the psychometric issues involved in the use of performance tasks can never be distilled down to a simple checklist, these guidelines are a useful starting point for more nuanced discussion. It is possible that machine scoring (automated scoring) may be implemented for some tasks (e.g., essays). As a result, we have included some criteria for evaluating both human and machine scores below. Any final decisions concerning performance assessment task acceptance will be made by the Consortium with the ultimate goal of verifying that each task would give the same result for the same student upon repeated administration.

While criteria for performance task acceptance can be set tentatively now, it is wise — especially with new assessments and new testing populations — to wait until actual student data are available to finalize these criteria. As tentative criteria, we recommend the following:

1. Absolute agreement of raters
   a. **Weighted kappa (using quadratic weights):**
      i. Rater-rater agreement must have a weighted kappa of .70 (rounded normally). This value was selected on the basis that it represents the “tipping point” at which signal outweighs noise in the prediction.
      ii. If condition (i) above is met, then the mean of weighted kappas for the autoscore-human rater 1 agreement and autoscore-human rater 2 agreement must also be .70 (rounded normally) for the automated scoring model to be accepted. That is, if the weighted kappa value for the automated score and human rater 1 is .68 and the value for the automated score and human rater 2 is .72, this would meet the criterion for use of the automated model.
   b. **Pearson correlations.** The mean of Pearson correlations, computed on rater 1 and rater 2 agreement, must also be .70 (rounded normally).

2. Degradation in agreement between rater and machine scores
   There can be no more than a .10 drop in agreement (measured as weighted kappa and/or correlation) from human-human agreement to automated-human agreement.
3. Mean difference

The standardized mean difference (standardized on the distribution of human scores) between human and automated scores cannot exceed .15.

Of these measures, two are preferred measures of reliability: Pearson product-moment correlations ($r_{ht}$) and (quadratic) weighted kappa statistics ($\kappa_{WGT}$). A further measure of association is the standardized mean difference ($\bar{z}$) between human or machine scores, which indicates whether scores are consistently higher or lower for one set of scores. The standardized mean difference can be computed as

$$\bar{z} = \frac{\bar{X}_{as} - \bar{X}_{H}}{\sqrt{SD^2_{as} + SD^2_{H}}},$$

where $\bar{X}_{as}$ is the mean of the automated score, $\bar{X}_{H}$ is the mean of the human score, $SD^2_{as}$ is the variance of the automated score, and $SD^2_{H}$ is the variance of the human score.

In addition, even in the case in which rater consistency studies indicate that sufficient reliability is demonstrated, there will likely still be large student-by-task interactions and contributions of other measurement facets to the overall variance. In that case, we would recommend conducting another generalizability study at the field-test stage.

2.10 Final Field Test Forms Verification

We will conduct analyses of simulated and operational field test adaptive administrations to monitor the functioning of the adaptive system and of each item pool used for operational field testing. We have a long history implementing, administering and monitoring item-level adaptive testing and will use this experience in working with the Consortium. From 1994 to 2011, we have administered nearly 15 million high-stakes item-level adaptive tests, and we have evaluated thousands of adaptive item pools. This experience has been invaluable in establishing various evaluation methods; the key to this and any evaluation is the set of criteria by which results are judged.

In order to evaluate the adaptive system, we will arrange to obtain all information needed to evaluate the delivery system (e.g., item response strings, resulting ability estimates produced, criteria, and specifications or constraints that the adaptive algorithm is attempting to meet), as well as the composition of each item pool used during field testing from the adaptive delivery Contractor. As stated in the RFP (p. 26), we will evaluate the following:

- fidelity of each summative test event, both real and simulated, to test blueprints and specifications
- measurement errors for simulated scores for both overall scores and subscores
Simulations will play an important role in evaluating the operational adaptive algorithm and delivery system. The simulations we conduct will be specific to the Consortium, using item parameter estimates from the field test items and simulated test taker populations that are representative of the students of the Consortium member states. We will work with the Consortium and its technical advisors to finalize the design of the simulation study, but as a starting point, we propose simulating 1,000 examinees at a given number of ability (theta) values (for example, 15) equally spaced between -3 and +3 and running each simulee through the adaptive algorithm. We can then summarize the results of those 15,000 test events and resulting ability estimates per item pool to examine the degree to which the algorithm and resulting scores meet the criteria outlined below.

While simulations are informative and convenient to conduct, they only provide one source of evaluation data. There is always a risk that the simulations may not adequately predict what happens when real students are administered operational tests. For that reason, wherever possible, we will evaluate results from real students tested during field testing, as well as from simulated data.

Fidelity of each summative test event to test blueprints and specifications. Early comparisons of adaptive testing procedures were made with regard to a narrow set of criteria. Foremost among these was test precision or its close associate, test efficiency. In a simulation, precision is defined as the degree with which the true underlying proficiencies of simulated test takers are recovered by simulated tests. Efficiency is simply test precision divided by test length. Precision and efficiency were highly prized because these are thought of as the principal “value-added” feature of adaptive testing. However, when a primary goal of testing is to find out what a student knows about a certain number of subscores within a content area, consistent content coverage is of equal or greater importance. A conforming test is one that meets all of the requirements imposed upon it. Conforming tests therefore both comply with all content constraints and measure to specified levels of precision. A better test administration algorithm is one capable of delivering a higher proportion of conforming tests.

To evaluate the fidelity or conformity of each test event to the test blueprints and specifications, for both simulated data and real test events, we will provide information about the content composition of the adaptive tests delivered from each item pool. During item selection, the algorithm is attempting to concurrently meet all specified criteria. We will create tables that summarize, for each criterion of the algorithm, both the mean number of items delivered and
the proportion of times each criterion is not met. These values will be reported for both the simulated and real data. The simulated data will provide a baseline for how we expect each item pool to perform. We would consider violations of constraints with higher weights/importance more serious than violations of constraints with lower weights.

**Measurement errors for simulated scores, both overall and subscores. — Test information functions. — Recovery of simulated examinee ability-analysis of bias and error.**

These three elements are highly interrelated, so they will be addressed together. The definition of test efficiency hinges on the corresponding definition of test precision. Test precision is loosely defined through the standard error of measurement. All test scores include an error component, the size of which generally varies across test takers. Differences in precision across score ranges are ignored by measures of precision that, like test reliability, are aggregated across score levels. However, IRT provides a related pair of test precision measures that are specific to or conditional on ability level. Both the test information function and the inversely related conditional standard error reflect the test precision level across the score scale. (The conditional standard error function is the inverse of the square root of the test information function.) In a simulation environment, the score bias function measures the extent to which score estimates converge to their true values. With smaller bias and error, test administration and scoring procedures recover simulated examinee ability better.

Even if the goal is to measure each student according to some fixed criteria for test information/conditional standard error, test precision can vary not just across proficiency levels but across test takers at the same proficiency level. Certain test takers are more easily measured compared with others. Many test takers respond predictably as the underlying item response model expects them to; for example, they correctly answer easy items and incorrectly answer difficult items. Such test takers will be more easily measured by an adaptive test than those test takers who respond in unexpected ways; for example, those that fail easy items and correctly answer difficult items. Predictable test takers can be well targeted from early in a test and presented a series of highly discriminating items (if an IRT model is used that measures item discrimination). Test takers who respond in unexpected ways will be more difficult to target and less likely to receive informative tests. Much of this inconsistency is unavoidable. However, test administration procedures may differ in the extent to which each test taker is measured according to the targeted precision level. It should be noted that exceeding the precision target is almost as undesirable as falling short. Measuring some test takers more precisely than necessary wastes resources — that is, increases item exposure — that could be used more productively with other test takers.

We will present the evaluation of how well the adaptive algorithm and item pool can recover simulated examinee ability by summarizing results for the 1,000 test events at each theta. For example, we will compute summary statistics for each 1,000 simulees with true overall and subscore thetas at given intervals. We will report conditional means, 25th percentiles, 75th percentiles, standard errors, and the proportion of times each criterion is not met. These values will be reported for both the simulated and real data. The simulated data will provide a baseline for how we expect each item pool to perform. We would consider violations of constraints with higher weights/importance more serious than violations of constraints with lower weights.
percentiles, conditional standard errors of measurement, and difference from target values for each theta interval. The conditional means and difference from target values will serve as indices of the ability of the algorithm and pool to recover the true abilities across the score range. The CSEM and 25th and 75th percentiles will serve as a measure of variability in reported scores for each true score.

**Analysis of summative pool adequacy for scores and subscores.** We will calculate a number of statistics to evaluate each of the summative pools used in field testing. Any given pool should be a compilation of all relevant item types representing all subscores, with varying levels of difficulty and item information. All pools used for any given test should be randomly equivalent. To investigate this, we will summarize the composition of each pool by reporting the number of various item types separately for each of the subscore levels. In addition, summary statistics of the IRT difficulty and discrimination parameters (if a 2PL or 3PL model is chosen) will be calculated for each pool and each subscore level. These results can be compared across summative pools to see if all pools are similarly composed. An easy way to see if overall pools or content subscores within pools are comparable against the average of all pools is to plot the pool and subscore information functions (please see Figure 7.a). Plots of the differences between each information function and the average can show where there is more or less information in an overall or subscore pool (please see Figure 7.b). For example, in Figure 7.b, the red pool (or subscore) has less information than the average pool in the region between thetas of zero and negative one, whereas the yellow pool has more information than the average pool around theta equal to 0.5.

**Figure 7.a: Plotting Sample No. 1.** Information function plots are a quick and easy way to check pool comparability.
Figure 7.b: Plotting Sample No. 2. Plots of difference values between each pool and the average of all pools can highlight areas where certain pools have more or less information.

ETS will also report expected and observed item exposure rates, where item exposure rate is defined as the proportion of the total number of examinees who were administered a particular item. We monitor item exposure for item and test security purposes to keep the same items from being administered to too many examinees, and also to keep pools viable by using as many items as possible. In pools with no exposure control, it is possible for 10 percent of the items to account for 70 percent to 80 percent of the items administered. We will calculate the frequency, percent, and cumulative percentage of items in each pool with various exposure rates. ETS will use simulated data to obtain the expected rates; we will use real data to obtain the observed rates. We will also include the correlation between expected and observed exposure rates, as well as summary statistics (mean, minimum, maximum, standard deviation) for exposure rate in this analysis. We will also examine overlap between simulated and operational adaptive tests. There will be less overlap with unconditional samples than samples conditioned on ability, so it is important to control (and monitor) exposure conditionally.

Test sustainability. This broad and vitally important criterion is not always considered in adaptive testing research. Essentially, sustainability refers to the ease with which a testing program can be operationally maintained over time. At least three factors are important here:

- What level of pretesting is needed to maintain summative bank stability? More sustainable testing programs will require less item development and pretesting to maintain summative test bank size and quality at stable levels.
- How balanced is summative test pool use? More sustainable testing programs will use items more effectively by balancing use. With balanced item use, every item appears with roughly equal frequency. When item use is poorly balanced, a
few items appear very often, and a large number are rarely or never used. Unbalanced item use affects sustainability by making a small number of exceptional items carry much of the burden. These items risk becoming known to the test-taker community and so may be retired from use, either temporarily or permanently. However, a large number of new items must be pretested to find the few that are exceptional enough to adequately replace those being retired. Under a more balanced test design, more commonplace items would be used often enough to reach retirement. Fewer new items would need to be pretested to replace these more typical items.

How easy are summative test pools to develop? Test administration procedures or algorithms that facilitate summative pool development will be more easily sustained over time. Several factors will influence the ease or difficulty of summative pool development, with some of these factors more easily quantified than others. One factor concerns the conditions that the pool must meet in order to be effective. Presumably, summative pools required to meet fewer and weaker conditions will be easier to develop. However, the extent to which pools parallel the structure of the summative bank is also important. Pools broadly representative of the summative bank will likely be easier to develop than pools that sample the bank more selectively. Finally, pools that operate in ways that are more predictable will be easier to develop than pools that function unpredictably. Minor changes to summative test pools should result in equally minor changes in the way a pool functions.

Ideally, test sustainability would be evaluated by simulations that predict the effects of several years of operational test administration. This simulation would start with the item banks as they currently stand and then work through several years of operational testing. Summative pools would be built, tests would be administered, item usage would be tracked, frequently administered items would be retired, and new items would be pretested and entered in the item bank. Comparing the summative bank at the end of this cycle with that at the outset would reveal whether the test administration procedures and all associated assumptions (item development requirements, pretest volumes, pool specifications, pool development, item retirement limits, etc.) were able to keep the item banks stable and the testing program sustainable over time.

Robustness. Test takers occasionally respond to test items in unexpected ways. Carelessness, speededness, item pre-exposure, unusual educational backgrounds, and a host of other factors can cause unexpected response patterns. Both conventional and adaptive tests are likely to poorly measure test takers who respond idiosyncratically. However, some adaptive administration and scoring procedures may cope better than others. We will conduct a series of simulations to evaluate the chosen procedures in this regard. Each simulation will be capable of generating data according to one of several identified nonstandard response models that would simulate the effects of careless responding, speededness, lucky guessing, and other sources of
anomalous responding. The evaluation will determine how successful the test administration and scoring procedures are in recovering true proficiency values despite the presence of unusual or aberrant responding.

2.11 Conduct Psychometric Analysis to Support Field Test Data Review

Field Test Item Analyses

In Section 2.7, we proposed a detailed set of analyses designed to evaluate the performance of test items, performance tasks, and forms for use following the pilot test. We propose that, in order to take advantage of the investment into the pilot test processes, systems, and report templates, we leverage as much as possible for the field test analyses, building upon and improving on this established foundation. As we noted in Sections 2.7, there is wide array of software available that might be used to meet the Consortium’s needs. The detail provided in Section 2.7 reflects the software packages that ETS believes could be useful to the Consortium, including our evaluation of how effective they are likely to be for Consortium work. Though we recommend those with an overall rank of one as being likeliest to support consistent, accurate results, we will work with the Consortium to create an effective solution with any of the software packages listed. ETS will make available any software code associated with any program as part of the technical documentation, as noted in Section 2.2. Section 2.7 provides detailed descriptions of all item and performance task analyses that we will perform to support the field test data review process. Below, we list and briefly describe the primary set of analyses and the activities we will complete in support of field test data review meetings and loading of data to the item bank.

1. Planned Field Test Analyses

a. Classical Item Analysis. In the first step for all analyses, ETS will include a classical item analysis to evaluate item difficulty, discrimination, and student raw score performance. We will use these analyses to identify any items that might not perform as expected and to determine the score reliability of the test form. We will flag items that do not meet the psychometric criteria decided upon by the Consortium.

b. Item Response Theory (IRT). IRT provides a unified statistical process for estimating stable characteristics of items and examinees and defining how these characteristics interact in describing item and test performance (Fitzpatrick et al., 2006). IRT item calibrations provide information that describes particular characteristics of individual items. The details of our approach in identifying the IRT model is described in Section 2.7, and the
analysis plan for calibrating the items and performance tasks included in the field test is described in Section 2.9.

c. **Differential Item Functioning (DIF).** ETS will also carry out DIF analyses for all items and performance tasks. DIF analyses are essential to helping verify that items are fair for all groups of students. As stated in Section 2.7, we will classify items into one of three categories based on threshold levels as noted. However, we are willing to work with the Consortium to develop the classification scheme that will work best for the Consortium.

d. **Accessibility Analyses for ELL and SWD.** We will conduct accessibility analyses for the field test phase similar to those described in Section 2.7 for the pilot. Analyses to be conducted for fixed and adaptive forms, when sample sizes allow, are (1) score reliability, (2) factor analysis, and (3) DIF.

ETS psychometric staff will work with the assessment development Contractor to carefully review all flagged items to determine the degree to which they are appropriate for use in the operational tests. We will refer items identified as not performing as expected based upon the full set of analyses to the Consortium, and these items may be revised and administered again at the Consortium’s discretion. We will provide items flagged for DIF to the Consortium, and also provide associated technical documentation.

a. **Performance Tasks.** ETS will produce data regarding the success of the performance tasks test versions for committee review, as noted in Section 2.7. ETS will produce for committee review indicators of the:

- consistency of human scores
- consistency of human scores and automated scores
- monitored real-time rater agreement
- categorical comparability across prompts and across years

In addition, we will present information about the statistical performance of remote (distributed) scoring systems for the committees’ and Consortium’s information.

b. **Adaptive Test Level Results.** ETS will produce test level results to support the analysis by demographic subgroup and for the adaptive test versions that have accessibility features. We will produce the analyses for adaptive forms as noted in Section 2.10, such that the appropriate Consortium committees have the necessary task data to evaluate the quality of the adaptive forms delivered and to evaluate that the adaptive system is working as intended.
2. **Data Review Meetings**

After the analysis of the field test data, we will work with the Consortium in preparing and conducting data review meetings. As described in Section 2.6, we recommend involving educators from across the Consortium in a thorough review of all pilot and field test items and performance tasks, including information on the performance of items under different conditions (e.g., paper and pencil, accommodated, and translated versions). To achieve this goal, we propose to conduct the field test data review in a similar manner as the pilot test data review (distributed system, online and/or face-to-face, and virtual meetings).

Following the training session, committee members will be asked to identify possible reasons for the results (e.g., whether they have detected a content problem within the item or task or whether there are instructional issues that may have negatively affected student performance). The committee will also be asked to recommend whether each flagged item and performance task is to be marked for revision, rejected, or accepted for inclusion in the Consortium item bank. In addition to the review of the item statistics, the committee members will also be asked to determine content appropriateness in light of empirical results. As with the pilot data review meetings, we expect to work closely with the Executive Director and Lead Psychometrician of the Consortium in planning and designing these meetings to maximize our opportunity to share and gather information with educators. Using the same or slightly modified criteria established for the pilot test data review, we will follow the same approach in preparing for these data reviews. The steps are described in detail in Section 2.6.

**Finalize and Upload Item Statistics**

After completing the analysis, and after resolution of the status of every item, all statistical information, annotations and educator feedback, and the final statistical status will be uploaded into the Consortium item bank. We will include all of the required variables within defined fields in the same upload file. The data format to be used has not yet been determined, but we expect to begin with industry standard approaches as we work with the item bank Contractor to produce the final format. We are experienced at providing data in multiple formats depending upon client needs or requirements.

Formats may vary in order to accommodate the specific test designs, but the general layout is fairly common. For example, we have experience with producing custom file formats when delivering information earmarked for external/non-ETS item banking systems. In most cases, file formats are derived from industry standard formats ranging from fixed length data records to QTI XML. A few of the file types we support at ETS include fixed length record, comma delimited (i.e., CSV, XLS, XML), and QTI XML. Regardless of the final agreed-upon file format, ETS is flexible enough to adapt to meet the input requirements of the Consortium item bank. Once the
contract is awarded, we expect to work closely with the Consortium and its item bank Contractor to understand the requirements and specifics of the item bank statistical upload process. At that time, we can determine the best approach and anchor the data exchange format.

2.12 Conduct Psychometric Analysis to Support Item and Task Calibration

1. Calibration/Linking Design. The details of the design, sampling, and calibration and linking plans for the field test are described in detail in Section 2.9. In that section, we proposed to use the method and IRT model that was identified during the pilot test stage; however, if our proposed hybrid design (described in Section 2.1.1) is accepted, even at the field test stage we will have the flexibility to modify the methods if necessary. That is, with the hybrid design we can calibrate and link the forms to the scales using both the common-item anchor (because the fixed linear forms and performance tasks will have overlapping content) and the equivalent groups design. As described previously, this calibration/linking design should be stronger than either common items or equivalent groups alone. However, if any part of the design cannot be met in the data collection phase, we can use one or the other method.

2. Calibration Study. To calibrate the field test data, we will use the IRT model and software chosen as a result of the studies completed with the pilot test data, as described in Section 2.7. If empirical evidence suggests that the choice of model should be revisited, then we will immediately notify the Consortium and prepare documentation for discussion with their TAC. As we described previously, we would use one of the commercially available software packages, acceptable to the Consortium to complete all of the analyses.

a. Missing data. Missing data can originate from a number of sources, either planned (by design) or unplanned, such as omitted responses or as a result of test speededness. Missing data “by design” is observed in block designs or vertical scaling designs (as we have described earlier), because students are only administered a portion of the items. Adaptive testing data, in particular, contains many occurrences of missing data where items simply were not administered to all examinees. Partitioning the larger sparse data matrix by grade level will help minimize the missing data in each response string; however, missing data will still be present “by design.” Most scaling programs have an option for indicating items that are missing by design, and the missing items will be treated as not reached and will be ignored during calibration.

Omitted responses that are unplanned can sometimes occur when a page layout is complicated, or following a long passage or task for which students fail to indicate a response due to fatigue or intimidation. Missing data can also occur where tests turn out to be speeded, and students run out of time. Other reasons for omitted responses include instances where students are
unmotivated, overly anxious, fatigued, or overwhelmed. When students omit responses to items, the items typically are scored as zero or not reached.

How to deal with unplanned omitted responses requires consideration of the underlying reason(s) for the omissions. The decision is a program policy one, because how the omitted items are scored can impact the student’s ability estimate or score. That is, scoring them as “not reached” will not impact a student’s score; however, scoring the omitted response as zero will lower a student’s ability estimate.

Another type of unplanned missing data can occur with performance tasks (constructed response) in which score levels are missing in the collected data — that is, when no student has achieved the highest score. While this is an unplanned source of missing data, it is something that can occur in new testing programs or with new item types, and may indicate a problem with the rubric or indicate that the item is too difficult. In this case, the general recommendation is to eliminate the item from the pool or fix the item in a later development phase and readminister the item at a future time.

b. Run time. Conducting the calibration studies should be relatively straightforward; however, we do recognize that large, sparse data matrices containing many items, as occur in vertical scaling, can present challenges to IRT estimation routines. Separate calibrations will greatly simplify estimation. Parameter estimates are usually achieved fairly quickly, with run times that take only minutes in many cases. Even for large, complex testing programs, several minutes are usually sufficient for convergence in the case of unidimensional models. If estimation fails to occur, then diagnosis and resolution of the problem will necessarily take longer. ETS will monitor and report all estimation diagnostics to the Consortium. In contrast to unidimensional models, multidimensional IRT models using Markov chain Monte Carlo (MCMC) sampling methods can require extensive time for estimation.

Even though the run times for the actual calibration are likely to be small, the processes that are required to be completed before the calibration study (e.g., data collection, processing, transferring, cleaning, and key verification) can take up most of the time in any schedule. We propose to minimize the processing and turnaround time by (1) conducting a thorough end-to-end test from the data collection through to the loading of data prior to the field test so that there will be no delay with any of the hand-offs, (2) automating as much of the processes as possible, (3) anticipating any potential issue or risk and developing a clear contingency or mitigation plan for each; and (4) having sufficient resources during the analysis phase to help assure that quality work is completed in the shortest time period. We are

We have conducted analyses and turned around results in short timeframes, and are confident we can meet SBAC’s timeframes.
experienced in conducting analyses and turning around results in timeframes as short as 48 hours, and we are confident that with the combination of the right processes with the right number and type of staff, we will meet the expected turnaround time.

c. Quality control procedures used in data processing. All psychometric analyses conducted at ETS undergo comprehensive quality checks by a team of psychometricians and data analysts. We complete detailed checklists to establish that each of the statistical procedures is performed correctly for every analysis. A group of senior psychometricians consisting of psychometric managers and directors will review the results of all calibrations.

During the item analysis, DIF, and calibration process, ETS checks that the correct options for the analyses are chosen. We also check the number of items, number of examinees with valid scores, IRT item parameter estimates, standard errors for the item parameter estimates, and the match of selected statistics to the results on the same statistics obtained during any preliminary item analyses. Our psychometricians also perform detailed reviews of item test plots and statistics to investigate whether the data fit the model. We check during the scaling process that the correct options for the analyses are used, and on the specifics of the scaling, such as the number of linking items, the average item difficulty for the linking items, the number of items dropped from the scaling, if any, and the scaling constants.

ETS psychometric and data analyst reviewers sign off on all checklists, the corresponding analyses, and output files. Our reviewers file electronic documentation that processes were precisely followed for each and every analysis.

In addition, we have extensive experience in documenting data processing procedures in enough detail for others to reproduce results. On several K–12 contracts, we have worked cooperatively with various independent Contractors and external auditors who parallel process our psychometric analyses to verify the results (e.g., AES, HUMRRO).

2.13 Present to TAC meetings as required

We understand that one of the important ways we must support the SBAC Consortium States is through helping them derive consensus on psychometric designs, issues, and decisions among Technical Advisory Committee (TAC) members, the Executive Committee, critical stakeholders such as the U. S. Department of Education, and the SBAC membership at large. Our program manager and the appropriate psychometric and research staff will work with SBAC to identify contract-related TAC meeting agenda topics in advance of the meeting and to prepare briefing materials to be sent to the TAC members at least 10 days prior to each meeting.

In addition, we will work with the Consortium staff to prepare and deliver meeting presentations, carefully listen to TAC discussions, and contribute to the discussions as needed.
and appropriate. If desired, our staff will take notes and provide these to the Consortium staff. We will send staff from the psychometrics and research areas that have a deep knowledge and understanding of the SBAC program, associated task work, and the issues at hand to each TAC meeting as required.

The current plan includes sending our Program Manager to each of the four annual meetings, and three members of our ETS Psychometric Team (Dr. Dianne Henderson-Montero, Dr. Richard Schwarz, and Dr. Wendy Yen) to each of the four annual meetings. In addition, our partners from Measured Progress (Dr. Michael Nering and Dr. Jennifer Dunn) are planning to attend three TAC meetings in Year 1, two TAC meetings in Year 2, and one TAC meeting in Year 3. Our experts will also be available through teleconferencing to participate in other meetings that may be required to reach consensus on psychometric topics, such as Executive Committee meetings, meetings with the U.S. Department of Education, and general SBAC membership meetings.

Should SBAC desire in-person representation at additional meetings, we will be happy to discuss this and determine how to accomplish it within the budget for the program. Members of the ETS Psychometric Services team are experienced in working with TACs for state programs and look forward to these opportunities to interact with the SBAC TAC and serve as information resources.
B. Work Plan
B. Work Plan

Introduction

The SBAC Psychometric Services Master Work Plan is large and complex, and it involves multiple vendors. We recognize and appreciate the complexity of the scaling, equating, and standard setting requirements. The combined psychometric teams from ETS and Measured Progress, along with expert consultants Dr. Wendy Yen and Dr. Daniel McCaffrey, will provide the Consortium with many of the leading minds in the industry.

If the ETS/Measured Progress Collaborative is chosen as the Psychometric Services Vendor, the SBAC requirements necessitate coordination among SBAC Leadership, Technical Advisory Committees, representatives from the Consortium States, and other SBAC contracted vendors for information that will inform this work—especially handoffs among various parties. We will work closely with the SBAC Leadership to define and monitor the various handoffs and coordination activities. We are prepared for intensive collaboration and tight schedules and are adept at responding quickly to new and emerging ideas.

Communication and collaboration are keys to the success of this project. As the Program Management section of this proposal details, the ETS Project Management team will maintain ongoing communication using a variety of tools so that all contributors are well-versed on all aspects of the project.

We have experience managing work plans that require multiple handoffs and collaboration among various contractors and stakeholders. ETS’s nine state contracts each involve at least one other contractor – Measured Progress in two of these instances. Accordingly, we realize that a program with SBAC’s comprehensive and collaborative vision demands a carefully thought out work plan and the flexibility to adjust the plan when circumstances merit modifications.

We look forward to working with the SBAC Leadership Team, staff, and committees; PARCC advisors; and current vendors by providing the plans for the complete Psychometric Services required for meeting the needs and expectations of the Consortium.

2.1 Determine Linking and Equating Design for Special Forms

2.1.1 Pilot Designs

The pilot stage is critical as it lays the foundation for the entire SBAC endeavor, well beyond the field test. We have carefully considered the wide variety of purposes the data collected are to serve, and in the Technical Requirements Section, present our design to fulfill the requirements for the necessary analyses. We will work with the Consortium, the TAC, and partners to refine the design so that all needs of SBAC are met. The plan allows for the evaluation of the performance characteristics of summative items and performance tasks, calibrating and linking
items to create both horizontal and vertical scales, examining the impact of test dimensionality on scores and proficiency scales, and evaluating the performance and comparability of both paper and accommodated test forms. We will work closely with the Consortium and the TAC to select the best sampling design such that the pilot sample represents the Consortium and can be adjusted accordingly to over-sample English learners, students requiring accommodations, and other demographic groups as necessary. This will provide a pilot sample that represents the Consortium in terms of key variables of interest and allows for the needed comparability studies. The proposed design takes into account both psychometric and practical considerations to provide sufficient data for all analyses needed to inform the field test and future of the SBAC.

We spent significant time developing an equating plan that capitalizes on the strengths of both common-item and randomly equivalent groups to create a strong linking for both horizontal and vertical scaling. Careful evaluation of analysis results is essential prior to, during, and after item response theory (IRT) scaling to examine the integrity of the vertical scales, as well as the quality of the equating of forms within grade. Prior to conducting the IRT vertical scaling, we will examine classical test statistics to evaluate the ordinality of linking items. During scaling, careful monitoring of the analyses is needed to examine aspects such as model-data fit, convergence criteria, and functioning of linking items. After IRT scaling we will compare the equated score distributions across grades, test characteristics curves (TCCs), and conditional standard errors (CSEMs), to see if the vertical scaling results appear reasonable. Similar analyses will take place to evaluate the horizontal equating to determine whether equated score distributions differ beyond what would be expected for sampling variation.

Finally, an important characteristic of the pilot test design is that the data collected must allow for the evaluation of the performance tasks and associated scoring processes. All of the data collection designs described in this proposal will yield sufficient numbers of examinees to complete the classical item analysis, differential item functioning (DIF), IRT analysis, rater analysis, and trend studies. We will work with the Consortium and partners to provide technical support for the scoring process.

2.1.2 Special (Fixed) Forms

The Consortium tests will need to include paper and pencil forms for students without access to electronic platforms, as well as special needs forms. For these forms we will establish form construction specifications that will satisfy the content coverage, cognitive complexity, and item type requirements for these fixed forms. We will complete a special data collection and study during the pilot test to place the paper and pencil and special needs forms onto the vertical scale. We will work with the Consortium to identify online forms that could serve as paper and pencil versions, and to design and implement an approach to link the forms to the vertical scale. We will conduct comparability analyses of the fixed forms administered during the pilot stage, and prepare a report for the Consortium for consideration. At the item level, analyses will include DIF analyses, scatterplots by mode of standardized classical item difficulties, and a
comparison of item parameter estimates and item fit statistics. At the test level we will make initial comparisons of score distributions in paper and online modes. We will also examine the summary statistics (mean, standard deviation, median), distribution of scores, the test characteristic curves and conditional standard error of measurement, as well as the reliability estimates of the forms under each of the conditions. We will also administer a computer familiarity questionnaire to both groups, including questions about computer access, use, and experience, as well as test preparation.

**2.1.3 The contractor will provide a detailed written description of all planned psychometric activities with sufficient time for SBAC staff and committees as appropriate to review and approve the plans.**

One of the keys to a successful project is clear communication and detailed planning of all psychometric activities. Due to the ground-breaking work of unprecedented scale that is being undertaken by the Consortium, we expect to work closely with not only the Consortium and the TAC, but with various other Contractors. Therefore, it is critical to clarify the nature of the plans, key deliverables, requirements from the Consortium and their other Contractors, and all handoffs among the vendors. Clearly written plans will help establish that all parties are informed of all activities. We will exercise flexibility in establishing sufficient time for the Consortium and its committees to review and provide input on plans and contract deliverables.

The twice-annual contractor meetings will be a key component of effective collaboration between ETS and the Consortium. These meetings will cover a broad scope of activities, including potential cross-Consortium Contractor activities. We will collaborate to plan these meetings to satisfy the needs of the Consortium, other collaborators, and the ETS/Measured Progress Collaborative. These meetings will allow the ETS/Measured Progress Collaborative team to review accomplishments, as well as candidly discuss future work plans and challenges, issues/concerns, schedules, commitments, Consortium expectations, and any required modifications to the plans and schedules.

**2.2 Develop Technical Manuals**

We recognize the need for delivering on-time and accurate SBAC Technical Manuals that will serve as the ultimate source of information about the SBAC assessment program, and they will be designed to serve the needs of various educators and stakeholders of the Consortium. We are dedicated to working with the Consortium and the Technical Advisory Committee to design comprehensive documentation that will provide evidence to support the validity of all aspects of the SBAC pilot and field tests. The Technical Manuals will include all aspects of the Consortium summative tests, including performance tasks, as well as information about the interim assessment components, from development and technical designs through implementation, item-, task-, and test-specific information corresponding to the pilot and field tests. In addition to including all of the required elements the Technical Manuals we will work to help establish
that the manuals meet all current and future USED peer review guidelines, contain detailed
descriptions of all procedures or analyses used to evaluate pilot and field test results and are
organized to permit easy cross-referencing to the Standards of Educational and Psychological
Testing (AERA, APA, & NCME, 1999). We envision one master document that is essentially a
compilation of various stand-alone technical reports to offer constituents flexibility to reference
a certain chapter of the Technical Manual without having to read the entire document. We also
propose creating an online version of the Technical Manuals in addition to any hard-copy or
electronic (PDF) versions and, should the Consortium be interested, we will work with them to
develop a technical documentation website, with hyperlinks to the Technical Manuals and
related documentation, as well as corresponding data access.

2.3 Determine Standard-Setting Design

ETS and Measured Progress have implemented standard-setting sessions in many states. Our
process for SBAC will be designed to not only draw on the collective experience of the
ETS/Measured Progress Collaborative, but will include an array of expertise from the larger
measurement community. Together, ETS, Measured Progress, SBAC, and the external
consultants will form the standard setting design committee (SSDC) for SBAC. The SSDC will
provide SBAC with a solution that is driven by a sound research-based foundation and not purely
by operational expertise. The SSDC will meet regularly so that operational design remains true
to research-based principles and meets SBAC needs and expectations. Throughout the contract
we will work with the TAC members in developing a solution that optimizes the balance
between client needs and psychometric rigor. We will present to the TAC with information
about the standard setting design so that their feedback and recommendations help establish
technical adequacy of the standard setting process. Upon SBAC approval, key PARCC
stakeholders will be invited to SSDC meetings in keeping with the collaborative spirit of the
education reform initiative, and provide the foundation for comparability across the two
consortia such that achievement level descriptors (ALDs) as both consortia are building their
assessments to measure the CCSS. Discussions between the two consortia should be ongoing
throughout the process, and we will work to facilitate these discussions. Finally, the SSDC will
serve to validate our design and recommend the evidence that should be collected to form and
support a comprehensive validity argument for the standard setting process.

2.4 Determine Vertical Scale Design

The Consortium’s plans call for vertical scales. We have experience creating vertical scales that
we can draw upon when developing the optimal data collection and sampling designs that will
allow for the construction of the SBAC vertical scale. This will require close collaboration with
the Consortium and the TAC to specify the number and types of scales required to best reflect
the underlying structure of the constructs, based upon the dimensionality study completed, and
to simultaneously provide scores that are both understandable and useful to the Consortium
and its stakeholders. Throughout this process, we expect to periodically involve the TAC in design decisions and interpretation of results, as determined by the Consortium.

2.5 Design Pilot Test Sampling Plan and Select Items and Tasks for Pilot Test Forms

Although the test blueprints for the Consortium’s assessment are not currently available, the goals and design of the pilot, in combination with the constraints for the pilot forms provide a framework that we can use to describe our proposed approach for pilot form construction. To meet the goals and timelines of the new assessment system we will work collaboratively with the Consortium and partners to meet all requirements of the pilot test. We will study the Consortium’s test specifications and blueprints for each content and grade to fully understand the expectations of how the developmental progression toward college and career readiness articulated in the Common Core State Standards will be operationalized in summative assessments and performance tasks for each grade level. Based on this review and in collaboration with the Consortium and the technical advisors and content experts, we will develop specifications to construct all pilot test forms. Because the statistical properties of the pilot items will be largely unknown prior to pilot testing, the selection of items to populate the pilot forms will require expert judgment, in relation to difficulty and expected performance of the items and their suitability for inclusion in the on-grade pilot form and in the vertical linking sets. We will work closely to support the assessment development Contractor and clarify forms constructions specifications for the test.

2.6 Develop Pilot Test Item and Task Data Review Materials

After pilot testing, it will be important to involve educators from across the Consortium in a thorough review of all pilot test items and performance tasks, including information on the performance of items under different conditions (e.g., paper and pencil and translated versions). An examination of the actual statistical results from the pilot test for all items and performance tasks will be an important part of this review. Clear communication and presentation of information is paramount to a successful data review meeting, but will be even more critical given the introduction of new item types, new delivery methods, and content associated with the new CCSS. The data and information must be presented in a clear and concise manner to maximize the opportunity for educators to discuss the observed results and provide feedback that can be used to prepare items for field testing. In the case of the performance tasks, there will need to be additional training for the educators attending the meetings with regard to evaluating the data from these items and the scoring methods and systems, as this may be unfamiliar to some attendees. We have extensive experience conducting both face-to-face and virtual meetings, and look forward to developing a system for data review that will meet the Consortium’s needs for educator engagement and efficiency in the review process. We expect to work closely with the Executive Director and Lead Psychometrician of the Consortium, in planning and designing
these meetings to maximize our opportunity to share and gather information with educators.

2.7 Conduct Psychometric Analysis to Support Pilot Test Data Review

Accurate evaluation of items, performance tasks, and the test forms in which they are represented will be the foundation of the SBAC assessment. To this end, we will employ careful and detailed attention when evaluating all analysis results, with the goal to identify any items or forms that might not perform as expected. Another crucial foundational decision is the selection of the IRT model that will be used to implement all scaling and equating work. We will work with the Consortium to evaluate and consider all available information that will lead to the best recommendation for the most appropriate IRT model (unidimensional or multidimensional) suited to the assessments and measurement goals of the Consortium. This decision must take into account the required test form information, the performance task scoring, accessibility for students with disabilities and English learners, and the comparability of paper and pencil and online versions. To inform our recommendation, we will complete a dimensionality study to inform the vertical scaling design, carefully considering the structure of the construct as represented by the data and use research supported evaluation criteria to support that recommendation. Finally, we will evaluate the implications for the measurement of growth.

Committee Review

We have extensive experience with leading committee reviews with the use of item cards. We recommend that the item card include the item itself and any information related to the item’s content on one side of the card and item statistics (classical and IRT) and differential item functioning (DIF) results on the other side of the. Information from the online/paper comparability study will appear as an additional DIF comparison. Our psychometricians will guide reviewers through the process with clear explanation in layman’s terms.

Software Use

A wide array of software is available that might be used to meet the Consortium’s needs. In this proposal we recommend several options for the Consortium to consider. We are also willing to work with other software packages that are not listed, as long as their psychometric properties have been established. We will supply the software code associated with any program as part of the technical documentation.

2.8 Conduct Analysis of Pilot Test

We will work with the assessment development, the delivery, and the scoring Contractors to produce a report summarizing recommended changes for the field test development process, based upon the pilot test data analyses. Thorough, careful evaluation of all pilot test results is crucial for providing recommendations for changes to improve the field test development
process. It will be particularly important to examine the implementation of innovative items, the performance of students in each content standard, and how students interacted with the technology associated with the delivery system. If our analysis determines that improvements could be made in any area of item development, we will work closely with the Consortium and the other Contractors to identify and make any potential improvements in the process.

**Analysis of anchor pools available for Field Testing**

As the SBAC assessment moves into the field test stage, evaluation of the psychometric properties of the anchor pools will be essential for the success of the forms developed for the field test. We will obtain content-balance evaluation of the pilot test anchor pools from the assessment development Contractor, and combine these results with our summary of the statistical properties of the anchor pools to produce field test forms. We will also base our recommendations for the field test development on traditional measures of reliability and validity used for fixed form tests, as well as test information functions and differential test function information, and the extent to which this statistical information matches the expected outcome for the pilot tests.

**Recommend changes to performance task scoring or data collection processes**

Based on the statistical properties of the pilot test performance tasks, and the feedback from experienced scoring experts on the kinds of responses generated by these tasks, we will provide formal recommendations regarding any proposed changes to scoring or data collection processes. We typically look at the distribution of item difficulty and the distribution of students across rubric score point values to help in determining whether the item pool has sufficient depth to cover future needs and to evaluate whether the rubrics appropriately delineate the range of student responses. Various types of statistics and review will help to identify problems. Our psychometric staff will take all of this information into account when discussing and providing recommendations for changes to the field test development process.

**2.9 Design Item and Student Sampling Plans for Field Test and Select Anchor Items and Tasks for Calibration and Building the Vertical Scale**

The SBAC program is technically complex and will have several Contractors. We understand that as the psychometric services partner, we will need to thoroughly understand each phase of the program’s development and that it will be essential to coordinate our efforts with the other Contractors to efficiently and effectively execute each phase of development in a manner that best meets the Consortium’s goals and objectives. To this end, ETS will work closely with all partners to document each phase of the project, identifying needed Contractor hand-offs and key deliverable deadlines to keep the project on track.
We understand that the field test will serve many purposes with the goal to refine the processes, procedures, and test characteristics in preparation for execution of the operational program. We will work closely with the Consortium and the TAC to implement data collection designs that yield data for all required analyses to implement the operational program. Additionally, we will work with the Consortium to design and test the online delivery platform for bugs, validate that the adaptive algorithm and item pools are working appropriately, and that the platform is functioning as intended. It will be critical that we work as active partners to anticipate and resolve any issues that are encountered, working with all other Contractors, work groups, the Consortium, and the TAC, to help establish that the test delivery at the field test stage adheres to the required design. We understand that this is a critical phase in the development of the SBAC assessment and we will advocate on behalf of psychometric considerations to arrive at the best possible solutions.

2.10 Final Field Test Forms Verification

Drawing upon our experience implementing, administering, and monitoring item-level adaptive testing, we will conduct analyses of simulated and operational field test adaptive administrations to monitor the functioning of the adaptive system and of each item pool used for operational field testing. We will work with the Consortium and the adaptive delivery Contractor to obtain the data necessary to allow us to evaluate the fidelity of each summative test event, both real and simulated. Our adaptive algorithm simulations will be specific; using item parameter estimates from the field test items and simulated test taker populations that are representative of the students of the Consortium member states. Wherever possible, we will evaluate results from real students tested during field testing, as well as from simulated data. We will create tables that summarize, for each criterion of the algorithm, both the mean number of items delivered and the proportion of times each criterion is not met. These values will be reported for both the simulated and real data. The simulated data will provide a baseline for how we expect each item pool to perform. We would consider violations of constraints with higher weights/importance more serious than violations of constraints with lower weights. ETS will also report expected and observed item exposure rates, and monitor item exposure for item and test security purposes, to avoid the same items being administered to too many examinees, and also to keep pools viable by utilizing as many items as possible. We will recommend procedures for comparing the summative bank at the end of an operational testing cycle with that at the outset of the operational program, to reveal whether the test administration procedures and all associated assumptions (item development requirements, pretest volumes, pool specifications, pool development, item retirement limits, etc.) were able to keep the item banks stable and the testing program sustainable over time.
2.11 Conduct Psychometric Analysis to Support Field Test Data Review

We proposed a detailed set of analyses designed to evaluate the performance of test items, performance tasks, and forms subsequent to the field test. We will take advantage of the investment into the pilot by leveraging, as much as possible, processes created under the pilot phase along with existing systems and report templates for completion of the field test analyses, building upon and improving this established foundation.

The primary set of analyses and the activities we will complete in support of field test data review meetings will include classical, IRT, and DIF analyses. We will work with the assessment development Contractor to carefully review all analyses and identify items not performing to standards acceptable for operational use. We will provide necessary documentation to the Consortium for the Data Review meetings and for upload into the Consortium item bank. In addition, we will present information about the statistical performance of remote (distributed) scoring systems.

We will produce test level analyses to support the analysis by demographic subgroups and for the adaptive test versions that have accessibility features. We will also complete analyses for adaptive forms such that the appropriate committees have the necessary task data to evaluate the quality of the adaptive forms delivered and to evaluate that the adaptive system is working as intended. To aid the Consortium and educators in a thorough review of all items and performance tasks, including information on the performance of items under different conditions, we will work with the Consortium and partners to improve (where applicable) the data review process implemented at the time of the pilot so the field test data review provides optimal information. We expect to work closely with the Consortium in planning and designing these meetings and educator training sessions, to maximize our opportunity to share and gather information. After the meetings, we will work with the Consortium to reconcile feedback and final decisions. After the data review process is complete, the final statistical status will be uploaded into the Consortium item bank using industry standard approaches to produce the final format.

2.12 Conduct Psychometric Analysis to Support Item and Task Calibration

We recognize that even at the field test stage we will need to exercise the flexibility to modify methods, if necessary. If empirical evidence suggests that the choice of model should be revisited, we will immediately notify the Consortium and prepare documentation for discussion with the TAC. This includes dealing with limitations of the data inherent with large data collection designs and various item types and delivery modes. Of particular interest at this stage is the estimation of final item parameter estimates and minimizing the time required to complete these analyses. In preparation for a calibration study to minimize the processing and
turn-around time, we will conduct a thorough, end-to-end test from the data collection through to the loading of data; automate as much of the processes as possible; work to anticipate any potential issue or risk and develop clear contingency or mitigation plans; and plan to have sufficient resources available during the analysis phase to establish that quality work is completed in the shortest time period. We are experienced in conducting analyses and turning around results in time frames as short as 48 hours, and are confident that, with the combination of the right processes and the right number and type of staff, we will meet the expected turnaround time. If estimation fails to occur, we will diligently diagnosis and work to resolve the problem monitoring all estimation diagnostics so they can be reported to the Consortium. On several K–12 contracts, we have worked cooperatively with various independent Contractors and external auditors who parallel process our psychometric analyses to verify the results, and we will document data processing procedures in enough detail for others to reproduce results.

2.13 Present to TAC meetings as required

We understand that one of the important ways we will support the Consortium States is through helping them derive consensus on psychometric designs, issues, and decisions among the TAC members, the Executive Committee, critical stakeholders, such as the U. S. Department of Education, and the SBAC membership at large. Our program manager and the appropriate psychometric and research staff will work with SBAC to identify contract-related TAC meeting agenda topics in advance of the meeting and prepare briefing materials. Additionally, we will work with the Consortium to prepare and deliver meeting presentations, carefully listen to TAC discussions, and contribute to the discussions as appropriate. If desired, our staff will take notes and provide these to the Consortium staff. Our experts will also be available through teleconferencing to participate in other meetings that may be required to reach consensus on psychometric topics, such as Executive Committee meetings, meetings with the U.S. Department of Education, and general SBAC membership meetings.
C. Project Schedule
C. Project Schedule

The inherent complexity of an overlapping schedule with complex dependencies, like those described by this RFP, requires excellent planning and management to help establish timely deliverables of the myriad tasks. We understand that the success of SBAC psychometric development activities, since it will impact other contracts and the final implementation, demands both thoughtful planning and flexibility.

We know that SBAC must coordinate activities across multiple contracts in order to create a consistent, coherent assessment system. Consequently, we will work diligently so that the schedule of activities for which ETS is responsible fits within the overall program schedule and that the appropriate coordination is in place to best enable our team to support the SBAC team in your efforts.

We have included a draft, high-level Microsoft Project Schedule of ETS-related activities and deliverables for the full duration of the contract. For this proposal narrative, the MS Project Schedule on the following pages confirms the RFP start and finish dates for each task and includes high-level subtasks and deliverables within each task.

Upon execution of the contract, we will work with the SBAC team to further expand the schedule and to create a baseline for discussion during the kick-off meeting.
<table>
<thead>
<tr>
<th>Task Name</th>
<th>Duration</th>
<th>Start</th>
<th>Finish</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBAC Psychometric Services Proposal Schedule</td>
<td>694 days</td>
<td>Mon 1/9/12</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td>SBAC Psychometric Services Begin</td>
<td>0 days</td>
<td>Mon 1/9/12</td>
<td>Mon 1/9/12</td>
</tr>
<tr>
<td>2.1 Determine Linking and Equating Design for Special Forms</td>
<td>694 days</td>
<td>Mon 1/9/12</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td>Psychometric team develops pilot design</td>
<td>10 days</td>
<td>Mon 1/9/12</td>
<td>Mon 1/23/12</td>
</tr>
<tr>
<td>Psychometric team develops pilot design accounting for special forms</td>
<td>10 days</td>
<td>Mon 1/9/12</td>
<td>Mon 1/23/12</td>
</tr>
<tr>
<td><strong>Deliverable A</strong>: Draft Plan for Linking and Equating Designs for Special Forms</td>
<td>0 days</td>
<td>Tue 2/14/12</td>
<td>Tue 2/14/12</td>
</tr>
<tr>
<td><strong>Deliverable B</strong>: Official Memorandum Agreed upon Design for Linking and Equating Special Forms</td>
<td>0 days</td>
<td>Wed 2/29/12</td>
<td>Wed 2/29/12</td>
</tr>
<tr>
<td>2.1.3 Psychometric Activities Plan</td>
<td>26 days</td>
<td>Tue 1/24/12</td>
<td>Wed 2/29/12</td>
</tr>
<tr>
<td>ETS writes draft of plan</td>
<td>5 days</td>
<td>Tue 1/24/12</td>
<td>Mon 1/30/12</td>
</tr>
<tr>
<td>Senior psychometrician reviews and provides technical comments</td>
<td>5 days</td>
<td>Tue 1/31/12</td>
<td>Mon 2/6/12</td>
</tr>
<tr>
<td>Revise and finalize plan</td>
<td>4 days</td>
<td>Tue 2/7/12</td>
<td>Fri 2/10/12</td>
</tr>
<tr>
<td>ETS sends plan to SBAC for review and approval</td>
<td>0 days</td>
<td>Fri 2/10/12</td>
<td>Fri 2/10/12</td>
</tr>
<tr>
<td>SBAC reviews and provides comments and/or signoff</td>
<td>10 days</td>
<td>Mon 2/13/12</td>
<td>Mon 2/27/12</td>
</tr>
<tr>
<td>Senior psychometrician incorporates SBAC comments into final</td>
<td>2 days</td>
<td>Tue 2/28/12</td>
<td>Wed 2/29/12</td>
</tr>
<tr>
<td>ETS Program management sends final approved plan to SBAC for record keeping purposes</td>
<td>0 days</td>
<td>Wed 2/29/12</td>
<td>Wed 2/29/12</td>
</tr>
<tr>
<td>2.1.4 Contractor Meetings</td>
<td>694 days</td>
<td>Mon 1/9/12</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td>Project face to face kick-off meeting</td>
<td>2 days</td>
<td>Mon 1/9/12</td>
<td>Tue 1/10/12</td>
</tr>
<tr>
<td>Attend twice annual contractor meetings</td>
<td>694 days</td>
<td>Mon 1/9/12</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td>2.2 Technical Manuals</td>
<td>694 days</td>
<td>Mon 1/9/12</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td>Senior psychometrician writes draft of technical manual</td>
<td>90 days</td>
<td>Mon 1/9/12</td>
<td>Tue 5/15/12</td>
</tr>
<tr>
<td>ETS sends technical manual to SBAC for review and approval</td>
<td>0 days</td>
<td>Tue 5/15/12</td>
<td>Tue 5/15/12</td>
</tr>
<tr>
<td>SBAC reviews and provides comments and/or signoff</td>
<td>50 days</td>
<td>Wed 5/16/12</td>
<td>Thu 7/26/12</td>
</tr>
<tr>
<td>Senior psychometrician incorporates SBAC comments into final version</td>
<td>5 days</td>
<td>Fri 7/27/12</td>
<td>Thu 8/2/12</td>
</tr>
<tr>
<td>ETS Program management sends final approved technical manual to SBAC for record keeping purposes</td>
<td>1 day</td>
<td>Fri 8/3/12</td>
<td>Fri 8/3/12</td>
</tr>
<tr>
<td>Develop Year Two and three technical manuals</td>
<td>509 days</td>
<td>Mon 10/1/12</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td><strong>Deliverable D</strong>: Final Comprehensive Technical Manual – Sept 30, 2014</td>
<td>0 days</td>
<td>Tue 9/30/14</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td>2.3 Determine Standard-Setting Design</td>
<td>525 days</td>
<td>Wed 2/1/12</td>
<td>Thu 2/27/14</td>
</tr>
<tr>
<td>Experts draft standard-setting design</td>
<td>20 days</td>
<td>Wed 2/1/12</td>
<td>Wed 2/29/12</td>
</tr>
<tr>
<td>ETS works with SBAC to revise and refine standard-setting design</td>
<td>490 days</td>
<td>Thu 3/1/12</td>
<td>Thu 2/6/14</td>
</tr>
<tr>
<td><strong>Deliverable E</strong>: Standard Setting Design</td>
<td>0 days</td>
<td>Wed 2/29/12</td>
<td>Wed 2/29/12</td>
</tr>
<tr>
<td>SBAC reviews and provides final comments and sign-off</td>
<td>15 days</td>
<td>Fri 2/7/14</td>
<td>Thu 2/27/14</td>
</tr>
<tr>
<td>Task Name</td>
<td>Duration</td>
<td>Start</td>
<td>Finish</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>-------------</td>
<td>--------------</td>
</tr>
<tr>
<td><strong>2.4 Determine Vertical Scale Design</strong></td>
<td>58 days</td>
<td>Mon 1/9/12</td>
<td>Fri 3/30/12</td>
</tr>
<tr>
<td>Psychometric team develops Vertical Scale Design</td>
<td>26 days</td>
<td>Mon 1/9/12</td>
<td>Tue 2/14/12</td>
</tr>
<tr>
<td>Senior psychometrician documents Design for SBAC review/input</td>
<td>12 days</td>
<td>Wed 2/15/12</td>
<td>Fri 3/2/12</td>
</tr>
<tr>
<td><strong>Deliverable F:</strong> Draft plan for Vertical Scale Design</td>
<td>0 days</td>
<td>Fri 3/2/12</td>
<td>Fri 3/2/12</td>
</tr>
<tr>
<td>SBAC reviews and provides comments and/or signoff</td>
<td>15 days</td>
<td>Mon 3/5/12</td>
<td>Fri 3/23/12</td>
</tr>
<tr>
<td>Senior psychometrician incorporates SBAC comments into final</td>
<td>5 days</td>
<td>Mon 3/26/12</td>
<td>Fri 3/30/12</td>
</tr>
<tr>
<td><strong>Deliverable G:</strong> Official Memorandum Agreed upon Design for the Vertical Scale Design</td>
<td>0 days</td>
<td>Fri 3/30/12</td>
<td>Fri 3/30/12</td>
</tr>
<tr>
<td><strong>2.5 Design Pilot Test Sampling Plan and Select Items and Tasks for Pilot Test Forms</strong></td>
<td>146 days</td>
<td>Wed 8/1/12</td>
<td>Fri 3/1/13</td>
</tr>
<tr>
<td>Psychometric Team develops pilot test sampling plan</td>
<td>20 days</td>
<td>Wed 8/1/12</td>
<td>Tue 8/28/12</td>
</tr>
<tr>
<td>Senior psychometrician documents pilot test sampling plan for SBAC review/input</td>
<td>7 days</td>
<td>Wed 8/29/12</td>
<td>Fri 9/7/12</td>
</tr>
<tr>
<td><strong>Deliverable H:</strong> Draft Plan for the Design of the Test Sampling Plan and Item Selection</td>
<td>1 day</td>
<td>Mon 9/10/12</td>
<td>Mon 9/10/12</td>
</tr>
<tr>
<td>SBAC reviews and provides comments and/or signoff</td>
<td>15 days</td>
<td>Tue 9/11/12</td>
<td>Mon 10/1/12</td>
</tr>
<tr>
<td>Senior psychometrician incorporates SBAC comments into final</td>
<td>5 days</td>
<td>Tue 10/2/12</td>
<td>Mon 10/8/12</td>
</tr>
<tr>
<td><strong>Deliverable I:</strong> Official Memorandum for the Design of the Test Sampling Plan and Item Selection</td>
<td>1 day</td>
<td>Tue 10/9/12</td>
<td>Tue 10/9/12</td>
</tr>
<tr>
<td>Assessment development contractor selects pilot items and tasks</td>
<td>30 days</td>
<td>Wed 10/10/12</td>
<td>Tue 11/20/12</td>
</tr>
<tr>
<td><strong>Deliverable J:</strong> Memorandum indicating psychometric final signoff on plan for selecting items and performance tasks for pilot forms</td>
<td>1 day</td>
<td>Wed 11/21/12</td>
<td>Wed 11/21/12</td>
</tr>
<tr>
<td>ETS reviews proposed pilot forms and summarizes recommended changes or provides signoff</td>
<td>3 days</td>
<td>Mon 11/26/12</td>
<td>Wed 11/28/12</td>
</tr>
<tr>
<td>ETS sends pilot form revision recommendations to assessment development contractor</td>
<td>1 day</td>
<td>Thu 11/29/12</td>
<td>Thu 11/29/12</td>
</tr>
<tr>
<td>Assessment development contractor revises and finalizes forms</td>
<td>25 days</td>
<td>Fri 11/30/12</td>
<td>Tue 1/8/13</td>
</tr>
<tr>
<td><strong>SBAC PILOT TEST ADMINISTRATION</strong></td>
<td>37 days</td>
<td>Wed 1/9/13</td>
<td>Fri 3/1/13</td>
</tr>
<tr>
<td>Begin administration</td>
<td>1 day</td>
<td>Wed 1/9/13</td>
<td>Wed 1/9/13</td>
</tr>
<tr>
<td>End administration</td>
<td>1 day</td>
<td>Fri 2/15/13</td>
<td>Fri 2/15/13</td>
</tr>
<tr>
<td>Scoring vendor scores performance task responses</td>
<td>10 days</td>
<td>Mon 2/18/13</td>
<td>Fri 3/1/13</td>
</tr>
<tr>
<td><strong>2.6 Develop Pilot Test Item and Task Data Review Materials</strong></td>
<td>293 days</td>
<td>Wed 2/15/12</td>
<td>Fri 4/12/13</td>
</tr>
<tr>
<td><strong>Deliverable K:</strong> Item Statistics File for Upload</td>
<td>0 days</td>
<td>Wed 2/15/12</td>
<td>Wed 2/15/12</td>
</tr>
<tr>
<td>Prepare materials for data review meeting</td>
<td>18 days</td>
<td>Tue 2/5/13</td>
<td>Thu 2/28/13</td>
</tr>
<tr>
<td><strong>Deliverable L:</strong> Data Review Decision Criteria</td>
<td>1 day</td>
<td>Wed 2/15/12</td>
<td>Wed 2/15/12</td>
</tr>
<tr>
<td>Send materials to SBAC for review and feedback</td>
<td>1 day</td>
<td>Fri 3/1/13</td>
<td>Fri 3/1/13</td>
</tr>
<tr>
<td>Train meeting participants and facilitate deliberations</td>
<td>20 days</td>
<td>Fri 3/1/13</td>
<td>Thu 3/28/13</td>
</tr>
<tr>
<td><strong>Deliverable M:</strong> Complete Item Statistics with Educator Feedback</td>
<td>10 days</td>
<td>Fri 3/29/13</td>
<td>Thu 4/11/13</td>
</tr>
<tr>
<td><strong>Deliverable N:</strong> Official Memorandum to comprehensively document consensus on all Data Review Materials and Processes</td>
<td>1 day</td>
<td>Fri 4/12/13</td>
<td>Fri 4/12/13</td>
</tr>
<tr>
<td>Task Name</td>
<td>Duration</td>
<td>Start</td>
<td>Finish</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>----------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td><strong>2.7 Conduct Psychometric Analysis to Support Pilot Test Data Review</strong></td>
<td>216 days</td>
<td>Thu 6/28/12</td>
<td>Tue 5/7/13</td>
</tr>
<tr>
<td>Begin Internal Preparation of Analysis Programs and file Layouts</td>
<td>140 days</td>
<td>Thu 6/28/12</td>
<td>Fri 1/18/13</td>
</tr>
<tr>
<td>ETS psychometricians write software requirements and analysis specifications</td>
<td>20 days</td>
<td>Thu 6/28/12</td>
<td>Thu 7/26/12</td>
</tr>
<tr>
<td>ETS programmers write, document and quality control check software programs to complete all analyses of pilot data</td>
<td>120 days</td>
<td>Fri 7/27/12</td>
<td>Fri 1/18/13</td>
</tr>
<tr>
<td>ETS data analysts work with Data Quality Services to define and document data layouts for all analysis data files</td>
<td>40 days</td>
<td>Fri 8/24/12</td>
<td>Fri 10/19/12</td>
</tr>
<tr>
<td>ETS Data Quality Services works with delivery contractors to define and document assessment data file layouts coming to ETS</td>
<td>40 days</td>
<td>Mon 10/22/12</td>
<td>Tue 12/18/12</td>
</tr>
<tr>
<td>End internal preparation of analysis programs and file layouts</td>
<td>0 days</td>
<td>Fri 1/18/13</td>
<td>Fri 1/18/13</td>
</tr>
<tr>
<td><strong>ETS complete statistical analysis of pilot data</strong></td>
<td>47 days</td>
<td>Fri 3/1/13</td>
<td>Tue 5/7/13</td>
</tr>
<tr>
<td>Begin statistical analysis of pilot data</td>
<td>0 days</td>
<td>Fri 3/1/13</td>
<td>Fri 3/1/13</td>
</tr>
<tr>
<td>ETS receives all data from delivery contractor(s)</td>
<td>1 day</td>
<td>Mon 3/4/13</td>
<td>Mon 3/4/13</td>
</tr>
<tr>
<td>ETS (merges data files - if needed) quality control checks that data file formats meet expected requirements</td>
<td>5 days</td>
<td>Tue 3/5/13</td>
<td>Mon 3/11/13</td>
</tr>
<tr>
<td>ETS data analysts complete analysis of pilot data</td>
<td>29 days</td>
<td>Tue 3/12/13</td>
<td>Fri 4/19/13</td>
</tr>
<tr>
<td>ETS psychometricians review analysis results</td>
<td>29 days</td>
<td>Mon 3/18/13</td>
<td>Thu 4/25/13</td>
</tr>
<tr>
<td>ETS data analysts produce item/task statistic offload files for SBAC item bank</td>
<td>2 days</td>
<td>Fri 4/26/13</td>
<td>Mon 4/29/13</td>
</tr>
<tr>
<td>ETS quality control checks SBAC item bank files and posts to secure site for SBAC</td>
<td>1 day</td>
<td>Tue 4/30/13</td>
<td>Tue 4/30/13</td>
</tr>
<tr>
<td>End Statistical Analysis of Pilot Data</td>
<td>0 days</td>
<td>Tue 4/30/13</td>
<td>Tue 4/30/13</td>
</tr>
<tr>
<td><strong>Deliverable O</strong>: Memorandum to report on the effectiveness, stability, and validity of the accessibility features</td>
<td>0 days</td>
<td>Tue 4/30/13</td>
<td>Tue 4/30/13</td>
</tr>
<tr>
<td><strong>Deliverable P</strong>: A document identifying the final choice of IRT model and a description of item and performance task data to be reported</td>
<td>0 days</td>
<td>Tue 4/30/13</td>
<td>Tue 4/30/13</td>
</tr>
<tr>
<td><strong>Deliverable Q</strong>: Comparability study report</td>
<td>0 days</td>
<td>Tue 4/30/13</td>
<td>Tue 4/30/13</td>
</tr>
<tr>
<td>SBAC Item Data Review Meetings</td>
<td>5 days</td>
<td>Wed 5/1/13</td>
<td>Tue 5/7/13</td>
</tr>
<tr>
<td><strong>2.8 Conduct Analysis of Pilot Test</strong></td>
<td>63 days</td>
<td>Fri 3/1/13</td>
<td>Thu 5/30/13</td>
</tr>
<tr>
<td>ETS receives summary input/recommendations from assessment development, data management and scoring contractors</td>
<td>0 days</td>
<td>Fri 3/1/13</td>
<td>Fri 3/1/13</td>
</tr>
<tr>
<td>ETS Senior psychometrician drafts report consolidating recommendations</td>
<td>53 days</td>
<td>Mon 3/4/13</td>
<td>Wed 5/15/13</td>
</tr>
<tr>
<td>Senior advisor(s) review and provide technical comments</td>
<td>5 days</td>
<td>Thu 5/16/13</td>
<td>Wed 5/22/13</td>
</tr>
<tr>
<td>ETS Senior psychometrician incorporates edits and finalizes report</td>
<td>5 days</td>
<td>Thu 5/23/13</td>
<td>Thu 5/30/13</td>
</tr>
<tr>
<td><strong>Deliverable R</strong>: A report detailing the analysis of the pilot test, including a discussion of the results</td>
<td>0 days</td>
<td>Thu 5/30/13</td>
<td>Thu 5/30/13</td>
</tr>
<tr>
<td><strong>2.9 Design Item and Student Sampling Plan and Select Anchor Items and Tasks for Calibration and Building the Vertical Scale</strong></td>
<td>190 days</td>
<td>Fri 5/31/13</td>
<td>Fri 2/28/14</td>
</tr>
<tr>
<td>ETS finalizes design for field test item and student sampling</td>
<td>15 days</td>
<td>Fri 5/31/13</td>
<td>Thu 6/20/13</td>
</tr>
<tr>
<td>ETS Client Management sends final deliverable to SBAC</td>
<td>1 day</td>
<td>Fri 6/21/13</td>
<td>Fri 6/21/13</td>
</tr>
<tr>
<td>ETS works with SBAC, TAC and AD contractor to set criteria for anchor items/tasks and anchor item/task sets and to determine decision-making criteria regarding review committee feedback</td>
<td>15 days</td>
<td>Mon 6/24/13</td>
<td>Mon 7/15/13</td>
</tr>
<tr>
<td>Task Name</td>
<td>Duration</td>
<td>Start</td>
<td>Finish</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>----------</td>
<td>-------------</td>
<td>--------------</td>
</tr>
<tr>
<td>ETS works with AD contractor and SBAC to establish composition of anchor item/task selection committee and review committee</td>
<td>20 days</td>
<td>Tue 7/16/13</td>
<td>Mon 8/12/13</td>
</tr>
<tr>
<td><strong>Deliverable S:</strong> Criteria for Selection of Anchor Items &amp; Committee Members</td>
<td>1 day</td>
<td>Tue 8/13/13</td>
<td>Tue 8/13/13</td>
</tr>
<tr>
<td>ETS recruits anchor item/task selection committee</td>
<td>40 days</td>
<td>Tue 8/13/13</td>
<td>Tue 10/8/13</td>
</tr>
<tr>
<td>ETS works with other contractors to implement final review/approval process by SBAC</td>
<td>98 days</td>
<td>Tue 7/16/13</td>
<td>Tue 12/3/13</td>
</tr>
<tr>
<td>Receive field test data from the scoring contractors after completion of field test administration</td>
<td>0 days</td>
<td>Tue 12/3/13</td>
<td>Tue 12/3/13</td>
</tr>
<tr>
<td>Calibrate new and revised items against the anchor items during the fixed-form phase of field testing</td>
<td>60 days</td>
<td>Wed 12/4/13</td>
<td>Thu 2/27/14</td>
</tr>
<tr>
<td>Complete final analyses for monitoring adaptive testing algorithm and performance of item bank</td>
<td>60 days</td>
<td>Wed 12/4/13</td>
<td>Thu 2/27/14</td>
</tr>
<tr>
<td><strong>Deliverable T:</strong> Written criteria for performance task acceptance</td>
<td>60 days</td>
<td>Wed 12/4/13</td>
<td>Thu 2/27/14</td>
</tr>
<tr>
<td>Recalibrate items on field test responses only as needed</td>
<td>60 days</td>
<td>Wed 12/4/13</td>
<td>Thu 2/27/14</td>
</tr>
<tr>
<td><strong>Deliverable U:</strong> A memorandum documenting key changes for the Field Test</td>
<td>1 day</td>
<td>Fri 2/28/14</td>
<td>Fri 2/28/14</td>
</tr>
<tr>
<td>2.10 Final Field Test Forms Verification</td>
<td>272 days</td>
<td>Fri 2/1/13</td>
<td>Wed 2/26/14</td>
</tr>
<tr>
<td>Senior psychometrician works with AD and adaptive delivery contractors to finalize field test form specifications</td>
<td>102 days</td>
<td>Fri 2/1/13</td>
<td>Tue 6/25/13</td>
</tr>
<tr>
<td>ETS will conduct analyses of both simulations and field test adaptive administrations to monitor the functioning of the adaptive system and item pool</td>
<td>130 days</td>
<td>Wed 6/26/13</td>
<td>Wed 1/1/14</td>
</tr>
<tr>
<td>ETS senior psychometrician drafts summary memorandum providing findings of the field test forms verification</td>
<td>30 days</td>
<td>Thu 1/2/14</td>
<td>Wed 1/2/14</td>
</tr>
<tr>
<td>Senior advisor(s) reviews and provides technical comments</td>
<td>5 days</td>
<td>Thu 2/13/14</td>
<td>Wed 2/19/14</td>
</tr>
<tr>
<td>ETS senior psychometrician finalizes memorandum</td>
<td>5 days</td>
<td>Thu 2/20/14</td>
<td>Wed 2/26/14</td>
</tr>
<tr>
<td><strong>Deliverable V:</strong> Results of the CAT simulations and field test administrations</td>
<td>0 days</td>
<td>Wed 2/26/14</td>
<td>Wed 2/26/14</td>
</tr>
<tr>
<td>2.11 Conduct Psychometric Analysis to Support Field Test Data Review</td>
<td>40 days</td>
<td>Thu 5/1/14</td>
<td>Thu 6/26/14</td>
</tr>
<tr>
<td>ETS complete statistical analysis of field test data</td>
<td>40 days</td>
<td>Thu 5/1/14</td>
<td>Thu 6/26/14</td>
</tr>
<tr>
<td>Begin statistical analysis of field test data</td>
<td>0 days</td>
<td>Thu 5/1/14</td>
<td>Thu 5/1/14</td>
</tr>
<tr>
<td>ETS receives all data from delivery contractor(s)</td>
<td>1 day</td>
<td>Fri 5/2/14</td>
<td>Fri 5/2/14</td>
</tr>
<tr>
<td>ETS (merges data files - if needed) quality control checks that data file formats meet expected requirements</td>
<td>5 days</td>
<td>Fri 5/2/14</td>
<td>Fri 5/9/14</td>
</tr>
<tr>
<td>ETS data analysts complete analysis of field test data</td>
<td>25 days</td>
<td>Mon 5/12/14</td>
<td>Fri 6/13/14</td>
</tr>
<tr>
<td>ETS psychometricians review analysis results</td>
<td>25 days</td>
<td>Thu 5/15/14</td>
<td>Wed 6/18/14</td>
</tr>
<tr>
<td>ETS data analysts produce item/task statistic offload files for SBAC item bank</td>
<td>5 days</td>
<td>Thu 6/19/14</td>
<td>Wed 6/25/14</td>
</tr>
<tr>
<td>End performs statistical analysis of field test data</td>
<td>0 days</td>
<td>Wed 6/25/14</td>
<td>Wed 6/25/14</td>
</tr>
<tr>
<td>ETS quality control checks SBAC item bank files and posts to secure site for SBAC</td>
<td>1 day</td>
<td>Thu 6/26/14</td>
<td>Thu 6/26/14</td>
</tr>
<tr>
<td><strong>Deliverable W:</strong> Data resulting from the field test analyses and a description of the methods used</td>
<td>0 days</td>
<td>Thu 6/26/14</td>
<td>Thu 6/26/14</td>
</tr>
<tr>
<td><strong>Deliverable X:</strong> Report of the Adaptive Test Results</td>
<td>0 days</td>
<td>Thu 6/26/14</td>
<td>Thu 6/26/14</td>
</tr>
<tr>
<td><strong>Deliverable Y:</strong> Recommendation for Data Storage Formats</td>
<td>0 days</td>
<td>Thu 6/26/14</td>
<td>Thu 6/26/14</td>
</tr>
<tr>
<td>Task Name</td>
<td>Duration</td>
<td>Start</td>
<td>Finish</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>----------</td>
<td>----------------</td>
<td>---------------</td>
</tr>
<tr>
<td><strong>2.12 Conduct Psychometric Analysis to Support Item and Task Calibration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ETS drafts plan for item calibration</td>
<td>8 days</td>
<td>Thu 6/26/14</td>
<td>Mon 7/7/14</td>
</tr>
<tr>
<td>Senior psychometrician reviews and provides technical comments</td>
<td>2 days</td>
<td>Tue 7/8/14</td>
<td>Wed 7/9/14</td>
</tr>
<tr>
<td>ETS senior psychometrician finalizes plan for calibrating item and performance tasks</td>
<td>2 days</td>
<td>Thu 7/10/14</td>
<td>Fri 7/11/14</td>
</tr>
<tr>
<td><strong>Deliverable Z</strong>: Report of data processing &amp; calibration procedures</td>
<td>1 day</td>
<td>Mon 7/14/14</td>
<td>Mon 7/14/14</td>
</tr>
<tr>
<td>SBAC provides feedback and/or signoff</td>
<td>5 days</td>
<td>Tue 7/15/14</td>
<td>Mon 7/21/14</td>
</tr>
<tr>
<td>Senior psychometrician incorporates SBAC comments into official memorandum</td>
<td>3 days</td>
<td>Tue 7/22/14</td>
<td>Thu 7/24/14</td>
</tr>
<tr>
<td>ETS provides official memorandum to SBAC</td>
<td>1 day</td>
<td>Fri 7/25/14</td>
<td>Fri 7/25/14</td>
</tr>
<tr>
<td><strong>2.13 Present to TAC Meeting</strong></td>
<td>694 days</td>
<td>Mon 1/9/12</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td>Prepare for quarterly TAC meetings</td>
<td>694 days</td>
<td>Mon 1/9/12</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td><strong>Deliverable AA</strong>: Presentations to SBAC Committee Meetings as required</td>
<td>694 days</td>
<td>Mon 1/9/12</td>
<td>Tue 9/30/14</td>
</tr>
<tr>
<td>Attend and present at other technical consensus meetings</td>
<td>694 days</td>
<td>Mon 1/9/12</td>
<td>Tue 9/30/14</td>
</tr>
</tbody>
</table>
Development schedules for the coming year should be presented at the annual planning meeting

We will submit a high-level draft schedule of Year 1 project activities, with separate timelines for each of the 13 tasks, to SBAC prior to the program kick-off meeting, with a detailed “Year 1 into Year 2” schedule to be completed and approved within one month of the meeting. Furthermore, we will submit an updated draft schedule of project activities, again with separate timelines for each of the 13 tasks, to SBAC a minimum of one month prior to each planning meeting.

At the initial kick-off meeting and yearly planning meetings, we expect that we will review these draft schedules with the SBAC in conjunction with the activities of its other contracts’ progress and work completed. We will revise the schedules to reflect a timeline that meets overall SBAC timelines and deadlines. We will submit the Microsoft Project Schedule to SBAC electronically, via the secure FTP site. Furthermore, we will work with SBAC closely on any required or requested schedule updates, and we acknowledge that all changes must be approved by the SBAC.

We use Microsoft Office Project 2010 and Microsoft Office Enterprise Project Management (EPM), a web-based software system that can be accessed and used by the SBAC team and contractor staff to securely access real-time project information and schedules. This tool helps us to establish that information is made available as needed, when needed, by all stakeholders. It includes an automated mechanism to issue email alerts when new information is posted to the project workspace. The Microsoft Project Schedule can also be submitted to the SBAC team electronically via email in another agreed format, if needed.

Part of the flexibility offered by the EPM system is the ability to create timelines and schedule views specific to any phase or deliverable of a project. As a result, ETS is able to provide easy-to-use, extracted timelines of the tasks, milestones, durations, and deadlines involved in each aspect of the 13 tasks. We will use these timelines to break down the comprehensive project schedule into manageable phases, without compromising the linkages and deadlines driving the full schedule.

Using the state-of-the-art EPM system also has proven to be valuable in management of our subcontractors and vendors.
D. Deliverables
D. Deliverables

The ETS/Measured Progress Collaborative will work collaboratively with the SBAC Leadership Team and, as needed, the SBAC work groups and Consortium representatives to establish a thorough review and consensus for each deliverable prior to finalization. We will give each deliverable a careful editorial review prior to finalization and posting.

During the initial planning meeting, ETS will work with SBAC to confirm our understanding of deliverables, determine acceptance criteria, identify delivery methods, and agree upon schedules and time-frames.

Where decisions and consensus are appropriate our proposed approach is to provide design documentation to SBAC that includes a series of submissions that allow for review and comment by appropriate SBAC representatives. First, we will submit a draft plan to SBAC and other stakeholders as required. The draft plan will include an appropriate level of detail so that all reviewers are fully aware of the proposed plan and methods. It will identify pending decision points and recommendations on specific tasks or deliverables. Following the outcome of those discussions, an interim document in the form of an interim official memorandum will contain all of the final decisions and recommendations made by the reviewers and will serve as the reference document and final scope of work. Once work is completed for each deliverable then this document, with any needed revisions, will contain all appropriate documentation of processes and deliverables and be included in the Technical Manuals.
Table D.1: Deliverable for Each Task. The following represent our current understanding of the deliverables associated with this project.

<table>
<thead>
<tr>
<th>Task</th>
<th>Task Name</th>
<th>Start</th>
<th>Finish</th>
<th>Deliverable Description</th>
</tr>
</thead>
</table>
**Deliverable B:** Official Memorandum Agreed upon Design for Linking and Equating Special Forms.  
ETS will provide draft plans to the Consortium with sufficient time for the Consortium and the TAC to review and provide input. After the lead psychometrician has incorporated SBAC’s input, ETS program management will deliver an official memorandum describing the agreed upon Linking and Equating Design for special forms to the Consortium. Content of this interim document will be included in the Technical Manual. |
At the conclusion of each year, ETS will deliver a formal technical manual documenting the psychometric services and products delivered. The final Technical Manual deliverable can be one comprehensive electronic report that is searchable. |
ETS/Measured Progress will provide a standard setting design document that describes the proposed standard setting methodology, including a process for selecting panelists, a process for review and approval by decision makers, and a procedure for revision and refinement of ALDs and benchmarking. |
**Deliverable G:** Official Memorandum Agreed upon Design for the Vertical Scale Design.  
ETS will provide draft plans to the Consortium with sufficient time for the Consortium and the TAC to review and provide input. After the lead psychometrician has incorporated the Consortium’s input, ETS program management... |
<table>
<thead>
<tr>
<th>Task</th>
<th>Task Name</th>
<th>Start</th>
<th>Finish</th>
<th>Deliverable Description</th>
</tr>
</thead>
</table>
| 2.5  | Design Pilot Test Sampling Plan and Select Items and Tasks for Pilot Test Forms | Aug. 2012 | Nov. 2012 | management will then deliver an official memorandum describing the vertical scaling design to the Consortium. Content of this interim document will be included in the Technical Manual.  
**Deliverable H:** Draft Plan for the Design of the Test Sampling Plan and Item Selection.  
**Deliverable I:** Official Memorandum for the Design of the Test Sampling Plan and Item Selection.  
ETS will provide a draft plan describing the process for linking, for fulfilling the test blueprint, and for the selection of the field test anchors and the interim assessment item bank. The plan will be provided to the Consortium with sufficient time for SBAC and the TAC to review and provide input. After the lead psychometrician has incorporated the Consortium’s input, ETS program management will then deliver an official memorandum describing the agreed upon pilot test sampling plan. Content of this interim document will be included in the Technical Manual.  
**Deliverable J:** Memorandum indicating psychometric final signoff on the plan for selecting items and performance tasks for pilot forms. |
ETS psychometric services will provide the item statistics files for upload into the Consortium’s item bank.  
**Deliverable L:** Data Review Decision Criteria.  
ETS will provide an official memorandum documenting established criteria for making decisions during data review of items and item/task writing improvements based on Pilot Test Data.  
**Deliverable M:** Complete Item Statistics with Educator Feedback.  
ETS psychometric services will provide the item statistics files for upload into the Consortium’s item bank with final record of statistical status and approved form of educator feedback to be stored for future use. |
<table>
<thead>
<tr>
<th>Task</th>
<th>Task Name</th>
<th>Start</th>
<th>Finish</th>
<th>Deliverable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Deliverable N:</strong> Final Official Memorandum to Comprehensively Document Consensus on all Data Review Materials and Processes. ETS will provide an official memorandum documenting reconciled feedback for final data review decisions made in consultation with the Consortium, a summary of final needs for item/task editing based on pilot review to prepare items for field testing, a summary of scorer agreement and reliability for performance tasks, and summary about the statistical performance of remote (distributed) scoring systems.</td>
</tr>
<tr>
<td>2.7</td>
<td>Conduct Psychometric Analysis to Support Pilot Test Data Review</td>
<td>Mar. 2013</td>
<td>Apr. 2013</td>
<td><strong>Deliverable O:</strong> Memorandum to Report on the Effectiveness, Stability, and Validity of the Accessibility Features. A memorandum reporting evidence that the accommodations and accessibility features are effective, do not alter the focal constructs, and provide valid assessment outcomes. Content of this document will be included in the Technical Manual. <strong>Deliverable P:</strong> A Document Identifying the Final Choice of IRT Model and a Description of Item and Performance Task Data to Be Reported. <strong>Deliverable Q:</strong> Comparability Study Report.</td>
</tr>
<tr>
<td>2.8</td>
<td>Conduct Analysis of Pilot Test</td>
<td>Mar. 2013</td>
<td>May 2013</td>
<td><strong>Deliverable R:</strong> A Report Detailing the Analysis of the Pilot Test, including a Discussion of the Results.</td>
</tr>
<tr>
<td>2.9</td>
<td>Design Item and Student Sampling Plan and Select Anchor Items and Tasks for Calibration and Building the Vertical Scale</td>
<td>May 2013</td>
<td>Feb. 2014</td>
<td><strong>Deliverable S:</strong> Criteria for Selection of Anchor Items and Committee Members. Written criteria and procedures for identifying anchor items/tasks, recruitment and selection of committees, and description of the review/approval process. <strong>Deliverable T:</strong> Written Criteria for Performance Task Acceptance. <strong>Deliverable U:</strong> A Memorandum Documenting Key Changes for the Field Test.</td>
</tr>
<tr>
<td>Task</td>
<td>Task Name</td>
<td>Start</td>
<td>Finish</td>
<td>Deliverable Description</td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------------------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2.10</td>
<td>Final Field Test Forms Verification</td>
<td>Feb. 2013</td>
<td>Feb. 2014</td>
<td><strong>Deliverable V:</strong> Results of the CAT Simulations and Field Test Administrations. An official memorandum describing the results of the simulations and field test adaptive administrations.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Deliverable V:</strong> Results of the CAT Simulations and Field Test Administrations. An official memorandum describing the results of the simulations and field test adaptive administrations.</td>
</tr>
<tr>
<td>2.11</td>
<td>Conduct Psychometric Analysis to Support Field Test Data Review</td>
<td>May 2014</td>
<td>Jun. 2014</td>
<td><strong>Deliverable W:</strong> Data Resulting from the Field Test Analyses and a Description of the Methods Used.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Deliverable W:</strong> Data Resulting from the Field Test Analyses and a Description of the Methods Used.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Deliverable X:</strong> Report of the Adaptive Test Results. An official memorandum describing the adaptive test results with specific attention to the performance tasks.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Deliverable X:</strong> Report of the Adaptive Test Results. An official memorandum describing the adaptive test results with specific attention to the performance tasks.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Deliverable Y:</strong> Recommendation for Data Storage Formats. A written proposal for formats for data storage in the Consortium item bank.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Deliverable Y:</strong> Recommendation for Data Storage Formats. A written proposal for formats for data storage in the Consortium item bank.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Deliverable Z:</strong> Report of Data Processing and Calibration Procedures. A document describing data processing and calibration procedures, including quality control processes and details to facilitate replication.</td>
</tr>
<tr>
<td>2.13</td>
<td>Present to TAC Meetings as Required</td>
<td>TBD</td>
<td></td>
<td><strong>Deliverable AA:</strong> Presentations to SBAC Committee Meetings, as required. Written documents addressing previously identified agenda items that can be distributed to meeting participants, as desired.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Deliverable AA:</strong> Presentations to SBAC Committee Meetings, as required. Written documents addressing previously identified agenda items that can be distributed to meeting participants, as desired.</td>
</tr>
</tbody>
</table>
E. Outcomes and Performance Measurement
E. Outcomes and Performance Measurement

Given the innovative nature of the SBAC assessment system, and the stakes that will be associated with results, SBAC is seeking a test designed to address issues and questions well in advance of operational implementation. To this end, the pilot and field test must be designed to meet the intended goals. As described throughout the RFP, the intended goals of the program are to:

» Evaluate item and task performance;
» Construct and evaluate the IRT vertical scaling consisting of multiple item formats;
» Construct and evaluate the provisional and final vertical scales;
» Identify the impact of construct dimensionality on test specifications, composite scores, and vertical scaling, and inferences concerning growth;
» Assemble provisional paper test forms and establish a concordance between linear computer based tests and provisional paper test forms;
» Evaluate test accommodations;
» Establish and evaluate materials and procedures for human scoring of constructed response tasks and ones pertaining to automated scoring approaches;
» Monitor psychometric characteristics (validity and reliability) of the adaptive testing algorithm;
» Produce periodic and comprehensive technical documentation;
» Support the growth model adopted and issues related to learning and instruction;
» Facilitate the development of item pool for interim assessments;
» Report and collaborate with various stakeholders such as data review (item writing), advisory committees, or Consortium members; and
» Plan, coordinate, and collaborate with other contractors.

The solutions provided in the Project Approach/Methodology Section recognize the importance of strong technical approaches and quality assessments necessary to achieve the Consortium’s objectives and goals. Our proposal describes technical approaches that will result in the highest degree of accuracy and evidence in support of validity, that are required for the purposes of measuring growth in student knowledge and skills toward college- and career-readiness in Mathematics and English Language Arts/Literacy. Psychometric services are a central component to the Consortium, and you need a psychometric contractor that has extensive experience in working with large, complex assessment programs with multiple vendors that are performing different functions and roles. Assessment design is often a series of trade-offs; ETS is
flexible and will provide a range of options with analyses that will facilitate decision-making on the part of the Consortium.

With such ambitious goals, a relatively short time-line, and multiple vendor involvement, it will be critical to have clear outcome measures at frequent intervals to document various design options, decisions, analyses, and results. The intent of the solutions provided in this proposal is to utilize sound methodology, present a range of different options, and demonstrate flexibility in the approaches suggested. Finally and importantly, there is the recognition that the psychometric results provided by a contractor must always be of the highest technical quality.

The quality of the work we do on the Psychometric Services project will be directly related to the quality of the summative assessment system that the SBAC states will use to test almost half the students in the country. State education agencies, boards, and legislatures will make important operational, policy, and funding decisions based on information from the SBAC assessments and the quality of those decisions rests heavily on the psychometric soundness of the program.

The ultimate outcome of our work will be its contribution to achieving the SBAC’s Theory of Action. We will not know that impact for years to come, which necessitates a series of interim outcomes and performance measurements that are indicative of the likelihood of achieving the ultimate goal.

In this section of our proposal, we propose a series of goals and targets, and accompanying measures of our performance, so that SBAC Leadership can monitor our performance on the contract. We present these in two major categories: Process Outcomes and Scope of Work Outcomes. Process Outcomes include things we will need to do that would increase the likelihood of success by establishing and maintaining an environment of working well with SBAC, the TAC and other vendors. They include collaboration tools, the project work plan, and communication methods, such as meetings and reports. These processes and tools defined below lead us to the Scope of Work Outcomes which include actual deliverable evidence that we have performed what we have been asked to perform for each of the thirteen main tasks in the Scope of Work

**Process Outcomes**

Process Outcomes are those deliverables that are meant to increase the likelihood of success in achieving the Scope of Work Outcomes, but also the quality of the working relationship between the ETS/Measured Progress Collaborative, and the other entities with whom we will work throughout the project.

**Collaboration Tools**
We realize that to be successful on this project, collaboration is going to be critical among the Consortium, the ETS/Measured Progress Collaborative, and other external vendors that will provide the required inputs from other SBAC projects.

In anticipation of the extremely iterative and collaborative nature of this work, ETS proposes to use a methodology for document sharing/revisions which will allow secure collaboration, sharing, and opportunities for review and subsequent modifications. This approach will allow SBAC and ETS to monitor progress on the deliverables.

**The Project Plan**

We use Microsoft Office Project 2010 and Microsoft Office Enterprise Project Management (EPM) tools. EPM is a web-based software system that can be accessed and used by SBAC and contractor staff to securely access real-time project information, schedules, and project documentation. This tool helps us establish that information is made available as needed, when needed, by all stakeholders. It includes an automated mechanism to issue email alerts when new information is posted to the project workspace, or when existing information has been changed. Utilizing the state-of-the-art EPM system has proven to be valuable in the management of our subcontractors and vendors, and allows us to be a better partner with our clients.

At the onset of the program, ETS will prepare a program plan built around the key activities and deliverables. This planning phase provides the fundamental guidelines for the remainder of the project. The remainder of the project requires the establishment of careful control and monitoring by the project management team, while our experienced group of researchers is working on the deliverables.

We will continually review each process and schedule for efficacy, work changes through with SBAC, and develop new Microsoft Project schedule plans as needed. The twice annual Contractor meetings and status meetings will focus heavily on such activities, though management meetings and weekly contact will further inform the process.

The project plan will delineate steps for all tasks and deliverables as part of the Psychometric Services project, including initiation and completion dates, as well as staff assignments. The ETS/Measured Progress Collaborative will work closely with SBAC to create the initial project plan. This plan will be reviewed and revised with SBAC periodically, with updated plans posted/delivered to SBAC.
Project plans will include details, dependencies, and subtasks for the following overall deliverables:

» Linking and Equating Design for Fixed Forms
» Technical Manuals
» Standard Setting Design
» Vertical Scale Design
» Item and Task Selection for Pilot Forms
» Pilot Test Items and Task Data Review Materials
» Pilot Test Analysis
» Selection of Anchor Items and Tasks for Calibration and Building Vertical Scale
» Field Test Forms Verification
» Field Test Data Review and Analysis
» Meeting Logistics (TAC meetings and Contractor meetings)

Communication Methods

At the onset of the project, the ETS Program Manager will work with the appropriate SBAC Leadership to define a structured communications plan for the program.

To allow the Consortium and our internal project teams to accurately monitor and evaluate the progress of the Psychometric Services project, the ETS project management team will maintain ongoing communications using a variety of methods and tools, including meetings, timely reports, and an escalation process.

Meetings

Proposed program meetings include:

» Program Kick-off and Planning meetings
» Semi-Annual Contractor meetings
» TAC meetings
» Internal Weekly Team meetings
» Other Ad-hoc Customer meetings, as needed (via web-ex, teleconference, etc.)

We will formally administer all meetings and sessions following a predefined agenda that will cover, at a minimum, project timelines, scheduled deliverables, risks, and the details of upcoming psychometric processes and analyses. Internally, Dr. James Augustin, Program Manager, will hold ongoing planned and “as needed” work sessions with team members so that
all contributors are well versed on all aspects of the project. These regularly-scheduled meetings will facilitate close collaboration with SBAC staff, which will help us to continually monitor the needs of the SBAC Psychometric Project. ETS will adhere to the SBAC requirements of posting and/or distributing meeting minutes within 3 days of meetings, monitoring and reviewing (and revising if needed) overall project plans.

**Status Reports**

Because of the complexity and critical nature of this work, ETS recognizes that more frequent status reporting will be required and we propose to provide weekly status reports to SBAC Leadership. We know that regular, meaningful, well-documented communication is a key foundational element of any successful project. One component of an effective communication strategy between SBAC and ETS will be our weekly status calls and subsequent reports. These scheduled weekly meetings will provide a regular opportunity to review the current schedule, discuss current deliverables, and update SBAC on ETS’s progress on upcoming tasks. The weekly management agenda and report will focus on:

» The Progress of Project Tasks  
» The Status of Deliverables  
» Issues and Concerns that May Arise during the Course of Contract Execution  
» Proposed Mitigation of Risks and an Analysis on Program Impact

**Escalation Process**

A critical component of the communications plan is an escalation strategy for notification regarding any issues that may arise during the program. This structure includes a plan for immediately communicating to the SBAC Director (or the appointed individual) via telephone, with follow-up in writing of any problem that has the potential to impact the quality, timeliness, or other aspect of the project, along with our proposed solution and a solution timeline. In addition, subsequent reports to SBAC will contain the issue, the determined solution, and current status within the solution timeline.

**Scope of Work Outcomes**

Whereas the outcomes described in the previous section will facilitate our performance on the project and allow the SBAC Leadership to monitor our performance through collaboration tools, a project plan and communication of progress on that plan, the ultimate outcomes are the deliverables for the thirteen major tasks in the Scope of Work.

All deliverables associated with the Scope of Work are listed in the project plan and will be incorporated into the detailed Microsoft Project schedule shared with the Consortium.
Several of the outcomes we intend to achieve are described below. Of course, we expect to finalize the list of outcomes jointly with the Consortium to make sure all necessary measures and outcomes are supplied to fully support the SBAC. As we wrote in Section 2.2 of the Technical Proposal, we intended much of the documentation to be used as stand-alone technical memorandums suitable for dissemination. These separate technical memorandums will also be included as chapters or sections within the overall Technical Manuals. Our Technical Manuals will fully document all aspects of the SBAC psychometric work such that a qualified independent contractor could replicate all aspects of the activities.

ETS Psychometric Outcomes:

» **Psychometric activities plan** for the pilot design, including all contractor hand-offs;

» **Vertical scaling planning document for both pilot and field test phases** that will include the decided upon innovative vertical linking design using common item scaling and flexible matrix design, the plan for administering sufficient numbers of items, and details on the plan for vertical scaling of performance task assessment components – the field test report will include revisions to the plan based on pilot test results;

» **Sampling plan document for both pilot and field test phases** that will record decisions about stratification sampling designs for obtaining representative samples that address the Consortium’s needs for participation (including which strata are chosen for stratified sampling and decisions made about use of statistical weighting), how certain populations will be oversampled to obtain sufficient data for analysis, as well as plans for forms construction, including how items and tasks will be distributed among the students sampled — the field test report will include revisions to the plan based on pilot test results;

» **Statistical analysis report for both pilot and field test phases** that will present all planned analyses in sufficient detail for an independent party to be able to replicate the results — including at least the following: analysis specifications, command files for the software used for analyses, and results of classical analyses including DIF and reliability, dimensionality analyses, IRT model selection analyses, performance task scoring inter-rater reliability results and generalizability analyses, and analyses of special fixed paper forms including accommodated and translated versions and associated comparability studies;

» **Data review results document for both pilot and field test phases** to record the processes used for data review and summaries of resulting decisions, as well as recommendations for incremental improvements in item and task development based on the statistics and/or committee suggestions;

» **Vertical scaling analysis report for both pilot and field test phases** that will describe comprehensive analysis and strategies for vertical scaling in terms of design, calibration approaches, IRT models used, dimensionality analysis (e.g., MIRT), and
scaling contingencies for both summative items and performance task assessment components — in the field-test version, cross-validation and approaches for “strengthening” the vertical scaling will also be included;

- **Standard setting report** that will describe designs considered that incorporate multiple approaches and sources of evidence, as well as all documentation of the meeting procedures and outcomes; and

- **Final field test forms validation report** that will include a description of how the algorithm, item pools, and all resulting test events were studied using both simulated and real data to monitor the functioning adaptive algorithms.

In addition to the outcomes listed above, all ongoing psychometric program decisions will be documented in writing and provided to the Consortium for your records. All pilot and field test item and task statistics and metadata will be provided to the item bank contractor. Options for software selection and final decisions made will be recorded, with additional provision for publishing any program code used.

Again, the list of all psychometric deliverables will be decided upon with the Consortium. Our goal is to provide you with the design, analyses, results and comprehensive documentation of all activities related to the SBAC. Precise chronicling of the pilot and field test phases will help establish successful operational administration of the SBAC assessments in the 2014-2015 school year.
F. Risks
F. Risks

The ETS/Measured Progress Collaborative believes that project risk is best handled by all team members including the SBAC contract manager by brainstorming and proactively identifying and categorizing the risks, ascertaining the impacts, and establishing both mitigation and contingency planning. A specific “owner” will be assigned to lead the resolution of each identified risk and a due date set for risk management plans to be in place. We will chart and review regularly Identified risks for updates and closure.

Risk Management

Our efforts to drive effective risk management throughout the program will be further supported through the use of Microsoft Office Enterprise Project Management (EPM), a web-based software system that SBAC and ETS staff can use to securely access real-time views of posted project information and schedules. By using a Microsoft Project schedule, we will be able to identify well in advance any possible delays and begin to discuss possible solutions to keep all deliverables on track.

Risks will be charted and monitored in the EPM system. Additionally EPM allows project personnel to view the updates to project risks in a real-time environment. This tool helps establish that information is made available as needed, when needed. Additionally, this tool has an automated mechanism to issue alerts when key messages require immediate attention. The EPM system also allows for the sharing of key project documents, current risks, and issues easily across the ETS team and with client staff in a secure web browser setting.

We will post weekly, user-friendly dashboard reports to communicate issues, risks, decisions, and the overall project status to stakeholders. Importantly, the EPM environment enhances our ability to determine project status at a glance. Consequently, it will enable the ETS Program Manager to create monthly briefing reports that provide the status of the Psychometric Services work in terms of schedule and key milestones, project scope, risks, and issues.

Business Continuity

Because we have many large, high-stakes testing programs throughout the world, ETS has also addressed some of the major business continuity risks faced in our highly technology driven world. ETS’s Business Continuity Management System (BCMS) is designed to safeguard our critical operations and systems, as well as the privacy and integrity of the information entrusted to us so that services essential to our customers are maintained in a time of crisis, and operations are resumed in a timely, orderly fashion.

BCMS is a leading-edge enterprise resiliency program, designed to identify risks, prioritize functions, and provide a detailed strategy for confronting natural and man-made emergencies. Based on best industry practices and standards, BCMS integrates more than 200 internal
business-continuity and disaster-recovery plans. Specifically, the ETS BCMS defines strategies and solutions that have been implemented and tested to recover and restore critical business processes, systems, and information required to continue operations in the event of a significant business interruption.

**Escalation Process**

A critical component of the communications plan is an escalation strategy for notification regarding any issues that may arise during the program. This structure includes a plan for immediately communicating to the SBAC Director (or the appointed individual) via telephone, with follow-up in writing of any problem that has the potential to impact the quality, timeliness, or other aspect of the project, along with our proposed solution and a solution timeline. In addition, subsequent reports to SBAC will contain the issue, the determined solution, and current status within the solution timeline.

**Risk Matrix**

As significant program-specific issues or risks arise, we will work in collaboration with the Consortium and TAC to effectively resolve them. We have listed several potential risks on the following pages along with a provisional solution approach for each. We propose this Risk Matrix format to drive discussion and resolution during program implementation, but will work with SBAC Leadership to determine the best format for presentation of risk information.
<table>
<thead>
<tr>
<th>Risk Title</th>
<th>Probability</th>
<th>Impact</th>
<th>Mitigation Plan</th>
<th>Contingency Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Setting procedures must be developed in concert with methods used by PARCC. PARCC may not be at the same stage of development as SBAC to allow coordination.</td>
<td>50%</td>
<td>SBAC procedures may not be informed by or inform the PARCC procedures.</td>
<td>Establish relationships with PARCC leadership to facilitate communication regarding drafts or access to appropriate PARCC TAC information.</td>
<td>Establish SBAC procedures and keep PARCC immediately informed for their reaction or to influence their work.</td>
</tr>
<tr>
<td>Final deliverables from other SBAC contract awards are not provided in a timely manner to meet the Psychometric Services schedules.</td>
<td>50%</td>
<td>Psychometric Services work may be pushed out due to incomplete information or data.</td>
<td>Establish a final schedule that aligns with current SBAC contract deliverables schedules.</td>
<td>Keep the SBAC contract manager informed of impending due dates for outside deliverables and adjust the psychometric work schedule if needed.</td>
</tr>
<tr>
<td>Open source software may not adequately serve the psychometric purposes intended.</td>
<td>25%</td>
<td>Search for better open source software or use 3rd party commercially available software if necessary.</td>
<td>Determine early limitations and availabilities.</td>
<td>Use the less preferred 3rd party commercially available software.</td>
</tr>
<tr>
<td>Pencil and paper administrations and mode comparability study occur in the pilot phase but significant item performance differences between pilot and field-test are evident.</td>
<td>25%</td>
<td>The final vertical scale will be adjusted at the field test.</td>
<td>The design in place for the Pilot Test must go through all necessary quality checks in order to obtain the best representative sample.</td>
<td>Consider readdressing the comparability study in the field test that also permits many more scaled field test items to be available for selection on subsequent paper forms.</td>
</tr>
<tr>
<td>Risk Title</td>
<td>Probability</td>
<td>Impact</td>
<td>Mitigation Plan</td>
<td>Contingency Plan</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>-------------</td>
<td>------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Consistent ability dimensions are evident in some grades but not consistently across the entire grade range. Unidimensionality may be supported across adjacent grades (e.g., 8 and 9) but not ranging from 3 to 11.</td>
<td>50%</td>
<td>The vertical scale articulation ranging from elementary and middle school levels with high school is not supported.</td>
<td>Determine based on the pilot phase which grades show a consistent pattern and where significant shifts in dimensionality occur. Create two vertical scales with a “bridge” grade overlapping in both scales (e.g., grade 8 in both lower and upper scales).</td>
<td>Use separate calibration methods and the bridge level to link the two scales in the field test.</td>
</tr>
<tr>
<td>Dimensionality results from the pilot indicate performance tasks follow a separate dimension from the summative items.</td>
<td>50%</td>
<td>Inability to build one scale to represent both the performance tasks and summative items.</td>
<td>The design in place for the Pilot Test must go through all necessary quality checks in order to obtain the best representative sample.</td>
<td>Compute a weighted score composite of separate summative test and performance tasks components. The weighting could either be statistical or derived from standard setting.</td>
</tr>
<tr>
<td>Student unfamiliarity with performance tasks results in omission, over estimates of difficulty and increased item attrition.</td>
<td>50%</td>
<td>Insufficient data exists to accurately scale the performance tasks.</td>
<td>Encourage SBAC to develop practice tests so that students and teachers gain increased familiarity with a range of performance task types and difficulty. The design in place for the Pilot Test must go through all necessary quality checks in order to obtain the best representative sample.</td>
<td>Retest the performance tasks and rely on field test data to build the scale.</td>
</tr>
</tbody>
</table>
3. Management Proposal
A. Project Management
Section 3 – Management Proposal

A. Project Management

In this section, we describe our proposed project management structure and staff and the qualifications of those individuals to provide superb service to the SBAC in this absolutely critical project. We have provided the Consortium with some of our most experienced people in every category. Given ETS’s deep bench of psychometric and technical staff, perhaps the largest in the country, we have selected individuals for their expertise and their track record of working collaboratively with clients and other vendors. Our subcontractor and collaborative partner, Measured Progress, has done the same.

Some of the work for this project will benefit from highly-specialized expertise to address certain challenges and issues that arise. In this case, we have named internal consultants, who are available to the project as needed. We have also established a system of executive oversight to make sure we are consistently delivering what SBAC needs. These senior executives are, in many cases, well known to the SBAC Leadership and your managing partner, and will be ready to handle any escalated issues to keep the project on track and your level of satisfaction consistently high.

1. Project Team Structure/Internal Controls

Project Management – Educational Testing Service

Our client-centered approach to providing excellent psychometric services and the related materials and documentation (the deliverables) begins with ETS program management and the ETS Psychometric Services Leadership Team. ETS proposes Dr. James Augustin as Program Manager. Dr. Augustin will have ultimate responsibility for all deliverables associated with the project. Supporting Dr. Augustin on the Leadership Team will be ETS Executive Program Director, Dr. Dirk Mattson, ETS Senior Psychometrician, Dr. Richard Schwarz, and Measured Progress Design Lead, Dr. Michael Nering.

» Dr. Augustin, Program Manager (ETS) – As the full time ETS program manager, James will provide the day to day program coordination as well as manage the wide-ranging details involved the schedule, risks, scope, and the deliverable work products for the program. He will serve as the principal point of contact between the SBAC and ETS including the subcontractors and consultants.

» Dr. Mattson, Executive Director (ETS) – Dirk will provide executive over site of program management functions. He will work with SBAC and the ETS team to build a relationship with SBAC staff that centers on cooperation and teamwork. His
experience in both the public and private sectors will enhance our understanding of your needs.

» Dr. Richard Schwarz, Senior Psychometrician (ETS) — Richard will serve as the lead psychometrician for the SBAC program. He will be accountable for the execution of all psychometric work and completion of deliverables under the psychometric services contract. He will lead work to incorporate input from both internal and external consultants, the Consortium, and the TAC.

» Dr. Michael Nering, Design Lead and Consultant (Measured Progress) — Michael serves as Assistant Vice President of the Research and Analysis division at Measured Progress. He is responsible for overseeing the daily operations of all psychometric, data analytic, and reporting services. Drawing upon years of psychometric experience, Dr. Nering provides psychometric support and initiates process improvement.

James Augustin and Dirk Mattson will focus primarily on program oversight and coordination with SBAC Leadership and stakeholders, while Rich Schwarz and Michael Nering will focus on establishing the best technical and psychometric quality of all designs, manuals and methodologies. To further provide the best designs possible, ETS will consult with highly-respected and experienced outside resources in the areas of standard setting, vertical scaling, and teacher performance.

SBAC will also have access to ETS senior management (as shown in the Organizational Chart) should issues need escalation or should additional consultation services be desired. SBAC should feel comfortable that your needs are being addressed by multiple groups and multiple levels at ETS.

The Measured Progress Psychometrics leadership/advisory team proposes to support ETS in standard setting design and other advisory roles for Psychometric Services, including contributions by Dr. Michael Nering, Dr. Jennifer Dunn, Dr. Michael Russell, and Dr. Luz Bay. The team structure is identified on the following pages:

**Organizational Structure – Educational Testing Service**

The project team will report through both James Augustin and Richard Schwarz who will have final authority over deliverables. Jim will serve as the Consortium main point of contact with ETS for this contract. Jim will work with Richard to lead the team in meeting the Consortium’s needs and delivering accurate, on-time materials.

The reporting lines for each staff member are indicated on the organizational chart on the following page. Staff who are not listed in a shaded box are not working directly on the program, but represent the management and reporting line for staff who are assigned to the Psychometric Services program team.
Project Management – Measured Progress

The Measured Progress Psychometrics leadership/advisory team proposed to support ETS in standard setting design and other advisory roles for SBAC-05 includes:

» Dr. Michael Nering, Design Lead and Consultant (Measured Progress) – Michael serves as Assistant Vice President of the Research and Analysis division. He is responsible for overseeing the daily operations of all psychometric, data analytic, and reporting services. Drawing upon years of psychometric experience, Dr. Nering provides psychometric support and initiates process improvement.

» Dr. Jennifer Dunn, Design Lead (Measured Progress) – Jennifer serves as Assistant Director of the Psychometrics and Research department. She is responsible for overseeing the daily operations of all psychometric functions. Drawing upon years of standard setting experience, Dr. Dunn has worked with clients across the country in establishing performance standards for a variety of situations and assessment types.

» Dr. Michael Russell, Consultant (Measured Progress) – Michael manages and leads research and development projects that focus on enhancing the validity of educational assessments through the use of digital technologies. He is trained in educational measurement and has eighteen years of experience conducting leading edge research and development of innovations to assessment.

» Dr. Luz Bay, Consultant (Measured Progress) – Luz is currently the Assistant Vice President within the Client Service division and oversees all NAEP standard setting contracts at Measured Progress. Through her previous roles at Measured Progress and other assessment companies, Dr. Bay has honed her analysis skills as well as developed methods to conduct quality assurance.

Collectively, this team has 40+ years of experience in psychometric study and assessment, have led dozens of research initiatives that focus on psychometric analysis, and published numerous papers, reports, and peer-reviewed articles on this topic. Finally, each team member has a history of leading collaborative projects that integrate the knowledge of state assessment program leaders, industry experts, and scholarly experts.

The proposed Measured Progress team to support ETS for the SBAC-05 program — both in leadership and support roles — has breadth and depth of experience and a strong track record of working with state assessment programs and national experts. They will work closely with ETS and SBAC to tap knowledge and expertise that will guide the development of future solutions that meet SBAC’s needs and to obtain SBAC feedback throughout the project.
**Organizational Structure – Measured Progress**

We propose structure for support of the Psychometric Services contract that features the team shown in the Organization Charts on the following pages. Measured Progress will serve in the following roles:

» Michael will serve as the Measured Progress design lead for standard setting design activities and as a consultant to ETS for other psychometric functions. Dr. Nering will have primary responsibility for the work conducted by Measured Progress. Dr. Dunn will be a design lead and will assist Dr. Nering with standard setting design support.

» Consultants — Dr. Michael Russell and Dr. Luz Bay will play a role in providing additional specific support to this work as indicated.

Project team members will report — for purposes of the work on each component — up through their organizational groupings (see organizational charts) to the project manager. Dr. Nering will serve as the Measured Progress primary point of contact for ETS and SBAC. He will also communicate, as needed and appropriate, with SBAC work group teams, members of the SBAC TAC, and SBAC members from all represented states.

**Organizational Charts and Table Indicating Lines of Authority**

As context, to explain to the SBAC and the proposal evaluators where proposed staff members fit within their corporate leadership structures, we have included a high-level corporate organization chart for ETS and Measured Progress. In addition, for each Measured Progress staff member named for the Psychometric Services program, we have added an accompanying table that indicates at least two levels of management above each team member.
Figure A.1: ETS Psychometric Services Program Team. The following chart illustrates the program team assigned to the SBAC Psychometric Services program.
Figure A.2: ETS Organizational Chart. The lines of authority for program staff are illustrated in this chart. The staff shown in gray are part of the Psychometric Services program team.
Figure A.3: Measured Progress Corporate Organizational Chart. The lines of authority for Measured Progress’ staff are illustrated in this chart.
**Internal Controls**

We will achieve success on the Psychometric Services project by use of effective project management practices and adherence to well-defined supporting processes and infrastructure.

We model our practices from the PMI’s *Project Management Body of Knowledge*, emphasizing the importance of planning, executing, and controlling the various aspects of a project. We understand the importance of effective project management, and we will apply these principles to the psychometric work to:

- Facilitate direct and open communication between the ETS project team, SBAC members, subcontractors, consultants, and other SBAC vendors. This can be accomplished by:
  - Meeting with the SBAC on a regular basis,
  - Making extensive use of teleconferencing for team meetings, and
  - Providing monthly briefing reports to SBAC that summarize the activities and deliverables

- Gain a thorough understanding of the project requirements (scope planning and definition) at the onset of the executed contract;

- Identify early the dependencies/inputs needed of this project on work being handled through other contracts in order to be able to minimize risks of delay in hand-offs;

- Use the requirements to develop a work plan and detailed project schedule(s) and manage project deliverables to their successful completion;

- Manage and control project scope and administer change control processes when appropriate;

- Document and resolve project issues and risks;

- Utilize the Enterprise Project Management (EPM) site for housing/maintaining project schedules, risk matrices, issues logs, documentation (to control versions), agenda, minutes, and all other relevant project information;

- Implement a structured communications strategy and leverage secure technology, where appropriate, to help facilitate quick and easy access to information and documents for project team members.

- Monitor budgets, invoicing, and resolve/provide further detail as requested.

These key areas of project management serve as the foundation for our best practices and are designed to support the on-time delivery of high-quality deliverables to our clients.
2. Staff Qualifications/Experience

ETS Staff Qualifications and Experience

Strong management and qualified staff are essential to completing the tasks and services for SBAC on time, within budget and with the highest quality. Our proposed SBAC team has the knowledge and experience to fulfill the requirements of the RFP. For project implementation, we will rely on proven processes and procedures that we have documented to be effective in similar projects. In addition, we are open to exploring innovative and cost-effective project improvements with SBAC. We believe that our many years of experience, combined with our desire to collaborate with and learn from SBAC, will allow our psychometricians to deliver superior services.

Our capabilities plus your expertise. Ultimately, our project management capability is vital to the success of every activity in every program that we manage. Noteworthy, too, is our recognition that each client we work with is unique in its requirements. Thus, we will look to the expertise of SBAC staff to guide us on the implementation of project objectives. Normally, we work with a single state and have access to that state’s expertise and that of its TAC. In this unprecedented project, we will have access to expertise from multiple states in the Consortium. We can already see from the quality of the Master Project Plan and the RFPs issued thus far that the caliber of thinking and planning is better than any we’ve seen in a single state.

Collaboration and flexibility. Our SBAC Psychometric Services team members will strive to build solid working relationships with SBAC committee members, work groups, leadership and other vendors. Our team will develop these relationships by listening and responding to the needs and requirements of both the Consortium States and the other SBAC contractors, demonstrating our reliability, and becoming immersed in the project.

Communicative rapport. The first and most important step that we take with each of our projects is to establish a communicative rapport with our clients. We understand that SBAC has an obligation to stakeholders — the Consortium States, and the students, parents, community members, teachers, and other educators within those states — and we take our commitment to helping meet that obligation seriously.

Highly qualified personnel. We have assembled a team of individuals who will work in collaboration with SBAC and stakeholders to manage a successful project designed to be of the highest possible quality. We chose our team with SBAC’s needs in mind.

ETS understands that SBAC has the right to approve any staff assigned to work on the Psychometric Services project and we will be happy to discuss any changes of staff that the SBAC requests throughout the contract period. Our goal is to provide the most qualified team of professionals who can work together with the SBAC Leadership to deliver the highest quality psychometric services. Should staff changes become necessary, whether originating at ETS or
through a request from SBAC, we will consult with SBAC as we provide additional staff recommendations and résumés for SBAC approval. We confirm that the staff members named in our proposal will perform the assigned work.

**Staffing Table**

The following staffing table lists all of the staff assigned to the program from both ETS and Measured Progress. The table lists the primary responsibility for each staff member, amount of time assigned to the program, and two levels of management. Staff biographies highlighting skills, education, experience and accomplishments for each person follow the staffing tables. Résumés for each staff member are provided in the *Appendix.*
<table>
<thead>
<tr>
<th>Function</th>
<th>Name and Title</th>
<th>Role in the Program</th>
<th>Percent Time on Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management</td>
<td><strong>Dirk Mattson</strong></td>
<td>Executive director provides program management oversight for all ETS contractual work completed for SBAC.</td>
<td>Year 1: 10% Year 2: 5% Year 3: 5%</td>
</tr>
<tr>
<td></td>
<td><strong>James Augustin</strong></td>
<td>Program manager serves as key client contact for the psychometric services contract and effectively manages all tasks that ETS is assigned to perform for the SBAC.</td>
<td>Year 1: 100% Year 2: 100% Year 3: 100%</td>
</tr>
<tr>
<td></td>
<td><strong>John Oswald</strong></td>
<td>Executive oversight of the psychometric services contract and program management team.</td>
<td>Year 1: 3% Year 2: 2% Year 3: 2%</td>
</tr>
<tr>
<td>Research and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychometrics</td>
<td><strong>Rich Schwarz</strong></td>
<td>Senior psychometrician leading and accountable for completion of all psychometric work for the contract.</td>
<td>Year 1: 100% Year 2: 100% Year 3: 100%</td>
</tr>
<tr>
<td></td>
<td><strong>Ying Lu</strong></td>
<td>Senior psychometrician assisting lead to complete all deliverables and execute all statistical analysis tasks for the project.</td>
<td>Year 1: 50% Year 2: 50% Year 3: 50%</td>
</tr>
<tr>
<td></td>
<td><strong>Bihua Xiang</strong></td>
<td>Establishes that the program team, internal and external consultants’ efforts are well coordinated.</td>
<td>Year 1: 50% Year 2: 50% Year 3: 50%</td>
</tr>
<tr>
<td></td>
<td><strong>Lin Lin</strong></td>
<td>Data Analyst leads completion of all analyst work of psychometric deliverables.</td>
<td>Year 1: 75% Year 2: 10% Year 3: 100%</td>
</tr>
<tr>
<td></td>
<td><strong>Manfred Steffen</strong></td>
<td>Consultation on evaluation of computer adaptive testing (CAT) item selection algorithms and monitoring use of CAT item pool.</td>
<td>Year 1: 5% Year 2: 5% Year 3: 5%</td>
</tr>
<tr>
<td></td>
<td><strong>Jeanette Graham</strong></td>
<td>Assists lead data analyst in completion of all statistical analyses for psychometric deliverables.</td>
<td>Year 1: 0% Year 2: 50% Year 3: 50%</td>
</tr>
<tr>
<td>Function</td>
<td>Name and Title</td>
<td>Role in the Program</td>
<td>Percent Time on Project</td>
</tr>
<tr>
<td>----------</td>
<td>---------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Year 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>01/09/12 - 9/30/12</td>
</tr>
</tbody>
</table>
| Research and Psychometrics | Jeannine Gilsdorf  
Manager, Data Analysis & Computational Research | Staff management only for overseeing programming work and transition of program code and program documentation to statistical analysis DA team for ongoing work. | 2%     | 2%     | 0%     |
|          | John Bonett  
Lead Research Systems Specialist | Expert Programmer to write and document all programming code for all analyses. | 30%    | 20%    | 0%     |
|          | John Cope  
Data Analyst Manager | Staff management only and overseeing work to transition programs from Data Analysis and computational research. | 2%     | 2%     | 0%     |
|          | Behroz Maneckshana  
Data Analysis Director | Accountable for data analysis oversight. | 5%     | 5%     | 5%     |
|          | Venessa Lall  
Senior Psychometric Director | Accountable for psychometric oversight. | 10%    | 10%    | 10%    |
|          | Dianne Henderson-Montero  
General Manager, Statistical Analysis, Principal Research for K-12, English Language Learners, and Global | Accountable for psychometric oversight. | 8%     | 10%    | 5%     |
|          | John Mazzeo  
Vice President, Statistical Analysis and Psychometric Research | Executive oversight of the psychometric services. | 1%     | 1%     | 1%     |
<table>
<thead>
<tr>
<th>Function</th>
<th>Name and Title</th>
<th>Role in the Program</th>
<th>Percent Time on Project</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Year 1 01/09/12 - 9/30/12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Senior psychometric advisor providing psychometric technical support for establishing that all documents and deliverables sent to the client and work completed are technically accurate.</td>
<td>7%</td>
</tr>
<tr>
<td>Psychometric Consultants</td>
<td>Jim Carlson Principal Psychometrician</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tim Davey Research Director</td>
<td>Senior psychometric advisor providing consultation for all technical decisions and analyses particularly as it pertains to computer adaptive testing.</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>Jonathan Weeks Associate Research Scientist</td>
<td>Consult on implementation of multidimensional item response theory (MIRT) and use of MIRT software along with Growth Modeling and analyses.</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Cara Laitusis Senior Research Scientist</td>
<td>Consultation on all issues to address inclusion and validity studies for Accommodations/Accessibility.</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Derrick Higgins Research Director</td>
<td>Consultation on evaluation of automated scoring technology.</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>Jay Breyer Principal Research Project Manager</td>
<td>Consultation on evaluation of automated scoring technology.</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>Chaitanya Ramineni Associate Principal Research Project Manager</td>
<td>Consultation on evaluation of automated scoring technology.</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Michael Kane Messick Chair Validity Research</td>
<td>Consultation on design implications for test score validity for all stakeholders.</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>John Young Research Director</td>
<td>Consultation on all issues to address inclusion and validity studies for English learners.</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Caroline Wylie Research Scientist Manager</td>
<td>Consultation on all issues to address formative/interim assessment implementation.</td>
<td>2%</td>
</tr>
<tr>
<td>Function</td>
<td>Name and Title</td>
<td>Role in the Program</td>
<td>Percent Time on Project</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Year 1</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>01/09/12 - 9/30/12</strong></td>
</tr>
<tr>
<td>Psychometric Consultants</td>
<td>Rick Tannenbaum</td>
<td>Provides consultation on standard setting issues as they pertain to psychometric services.</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Research Director</td>
<td></td>
<td><strong>Year 2</strong></td>
</tr>
<tr>
<td></td>
<td>Patricia Baron</td>
<td>Provides consultation on standard setting issues as they pertain to psychometric services.</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>Senior Research Project Manager</td>
<td></td>
<td><strong>Year 3</strong></td>
</tr>
<tr>
<td></td>
<td>Wendy Yen</td>
<td>Senior Psychometric Advisor, providing overall psychometric consultation particularly for vertical scaling, linking/equating; consult on preparation for and attend at technical advisory meetings.</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td>Consultant</td>
<td></td>
<td><strong>Year 3</strong></td>
</tr>
<tr>
<td></td>
<td>Daniel McCaffrey</td>
<td>Provides consultation on implementation and use of growth models for SBAC.</td>
<td>3%</td>
</tr>
<tr>
<td>Measured Progress</td>
<td>Michael Nering</td>
<td>Providing consultation on standard setting and other issues as they pertain to psychometric services.</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Associate VP Research and Analysis</td>
<td></td>
<td><strong>Year 3</strong></td>
</tr>
<tr>
<td></td>
<td>Role: Standard Setting</td>
<td></td>
<td><strong>Year 3</strong></td>
</tr>
<tr>
<td></td>
<td>Jennifer Dunn</td>
<td>Providing consultation on standard setting as they pertain to psychometric services.</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Assistant Director, Psychometrics</td>
<td></td>
<td><strong>Year 3</strong></td>
</tr>
<tr>
<td></td>
<td>Luz Bay</td>
<td>Providing consultation on standard setting as they pertain to psychometric services.</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>Associate VP Client Services</td>
<td></td>
<td><strong>Year 3</strong></td>
</tr>
<tr>
<td></td>
<td>Michael Russell</td>
<td>Providing consultation on accessibility issues as they pertain to psychometric services.</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>Vice President Nimble Innovation Lab</td>
<td></td>
<td><strong>Year 3</strong></td>
</tr>
</tbody>
</table>
Staff Biographies for the Psychometric Services Program Team

Our staffing tables for the Psychometric Services contract show all named staff from ETS and Measured Progress, their primary responsibility (role) for this contract, and the amount of time each will be assigned to this project for the specified contract period.

Staff biographies are presented in alphabetical order by category as presented in the organizational chart.

ETS Management Team

Dr. Dirk Mattson, ETS Executive Director, SBAC, is joining ETS to oversee every aspect of work for SBAC. He comes to us from Measured Progress, where was a Vice President responsible for leading the development of next generation assessment solutions, including APIP application, item innovation, and performance assessments. Previously, Dr. Mattson worked for eight years at the Minnesota Department of Education (MDE). From 2006-2011 he was the MDE’s Director of Research & Assessment, providing psychometric, policy, and management-level leadership for Minnesota’s statewide assessment program. He was responsible for supervising the development and administration the tests, implementing state policies, and supervising the creation of valid and reliable measurements of student achievement and system accountability. From 2003-2006, he was the MDE’s manager of Test Development, overseeing item development and test construction for the Minnesota Comprehensive Assessments-Series II. He has also worked as a public school curriculum coordinator, special programs coordinator, and classroom teacher. Dr. Mattson is affiliated with the American Educational Research Association and National Association of Test Directors. He earned his Ph.D. in Curriculum & Instruction from Arizona State University, his M.S. in English Education from Minnesota State University, and his B.A. in Secondary Education (English & Social Studies) from Concordia College in St. Paul, MN.

Dr. James Augustin, ETS Program Manager, coordinates ETS program activities and establishes that all contract deliverables meet requirements and are complete. In addition to working as a contractor for several assessment companies (including ETS) in recent years, he previously served as the Director of Large-Scale Assessment at ETS. As the director he managed all development activities for several state testing programs, including Georgia, Indiana, Maryland, and New Jersey, as well as the Commonwealth of Puerto Rico. In this role Dr. Augustin was the point of contact between clients and ETS staff, he prepared plans and schedules against contractual requirements, and he monitored team progress toward the completion of contract deliverables. He also was involved with standard setting activities for the Georgia, Indiana, and Puerto Rico programs. Prior to being hired at ETS, He coordinated the development of educational assessment programs for large-scale clients at The Psychological Corporation/Harcourt Education Measurement. Dr. Augustin has published numerous articles in psychology and education journals. In 2002 he was a guest editor of a special issue of Measurement and Evaluation in Counseling and Development on topics related to large-scale, high stakes assessment in K–12 education settings. He earned his Ph.D. in Psychology from
North Carolina State University, his M.A. in Psychology from Marquette University, Milwaukee, WI, and his B.A. in Psychology Trinity University, San Antonio, TX.

**ETS Psychometric Services Team**

**John Mazzeo, Vice President, Statistical Analysis, Data Analysis & Psychometric Research, Research & Development**, is responsible for the overall leadership of these three areas that provide the following: statistical and psychometric support for ETS testing programs; data analysis services and statistical programming support for various R&D activities; and foundational statistical and psychometric research, as well as statistical consulting in support of R&D and other ETS departments.

Mazzeo’s tenure with ETS spans more than 25 years in numerous departments, most within R&D. He joined ETS in 1984 as an Associate Measurement Statistician for The College Board® Statistical Analysis department. Subsequent positions within the organization included Director of State National Association of Educational Progress (NAEP) Analysis (1990–1994); Executive Director and Associate Project Director for NAEP (1994–2000); Center Director and Associate Project Director for NAEP in the Large-Scale Assessment Research Center (2000–2002); and Senior Research Director for Statistical and Psychometric Research and Deputy Project Director for NAEP (2002–2004). Mazzeo was promoted to Associate Vice President for Statistics and Psychometrics in 2004 and then to Vice President in 2009.

He is the author of numerous journal articles as well as research reports on assessment and data analysis and has presented more than three dozen papers at national and international conferences for the education community. Mazzeo also is a member of the American Educational Research Association, the National Council on Measurement in Education and the Northeastern Educational Research Association. He earned his bachelor’s degree in psychology from the State University of New York at Binghamton, and both a master of education in educational psychology as well as a doctorate in educational measurement and statistics from the University of Arizona.

**Dr. Dianne Henderson-Montero, General Manager and ETS Senior Advisor**, will serve as senior advisor for SBAC. She is the General Manager for ETS Statistical Analysis & Research, managing a staff of approximately 80 psychometricians, researchers, and data analysts, who work on the research designs of large-scale, complex testing programs. Joining ETS in 2001 after two years as a research scientist at CTB/McGraw-Hill, Dr. Henderson-Montero has more than 12 years of experience in the educational testing field and has held positions of increasing responsibility. Her experience in the field ranges from large district assessment and accountability programs, state testing programs, global in-country assessments of student performance, and assessments of English language, both domestically and internationally. Currently, her responsibilities include supervising, monitoring, and providing psychometric oversight and expert guidance for complex assessment programs to verify compliance with client specifications, ETS policies, and sound
measurement principles for several large-scale custom assessment programs as well as products for the assessment of English language proficiency. Her exemplar programs include:

» National Assessment of Educational Progress (NAEP)
» California Standardized Testing and Reporting (STAR)
» California High School Exit Examination (CAHSEE)
» California Alternate Performance Assessment (CAPA)
» Maryland High School Assessment (MD HSA)
» Tennessee End of Course Assessment
» Test of English as a Foreign Language (TOEFL)
» Test of English for International Communication (TOEIC)
» Washington Comprehensive Assessment Program (WCAP)

She is an expert in many aspects of educational design and measurement, including test design, psychometrics, equating, and the end-to-end process for item development, test scoring methods and score reporting. Dianne is characterized as a creative leader and strategic planner who is expert in enhancing communication, knowledge transfer, and fostering collaboration among staff across departments to develop creative and innovative solutions. Skilled in capacity and demand management, process improvement, as well as budgeting, forecasting, and operational process management, she has implemented key systems to facilitate the smooth operation of the organization, decrease costs, mitigate risk, and communicate status on all projects in her portfolio. Dianne earned a Ph.D. in Educational Psychology from the University of Alberta and is affiliated with the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME).

Dr. James Carlson, ETS Principal Psychometrician, will assists in the design and statistical analyses of the SBAC. He brings more than 30 years of research and psychometrics experience to K–12. Prior to rejoining ETS in 2005, he served as Assistant Director for Psychometrics at the National Assessment Governing Board, responsible for policy decisions regarding the National Assessment of Educational Progress (NAEP). He also worked at CTB/McGraw-Hill and ETS as chief research scientist and senior research scientist/director of research, respectively. Dr. Carlson cultivated knowledge of classroom teaching while leading computing services and research at Auburn University; directing program support and research at American College Testing, now ACT; and teaching graduate courses in statistics and computing at the University of Ottawa (Canada) and the University of Pittsburgh. In addition to his current work for ETS, he currently also edits the ETS Research Report Series, and previously edited the Journal of Educational Measurement. He earned his degrees at the University of Alberta (Canada): a Ph.D. and M.A. in Educational Psychology and a B.Ed. in Mathematics Education.
Dr. Tim Davey, ETS Research Director, will bring his statistical experience to the SBAC. He joined ETS in 2000, and oversees work at the Center for Psychometric Infrastructure. He previously served as Executive Director of the for Graduate and Professional examinations. He is interested in developing and improving procedures for administering and supporting computerized and adaptive tests. At American College Testing (ACT), where he worked for 12 years, Dr. Davey was responsible for directing development of computerized and adaptive testing programs, leading project teams conducting and presenting research in psychometrics and measurement, supervising computer programmers and support staff, and developing statistical and measurement procedures. He earned his Ph.D. in Psychology with a concentration in Measurement, an M.S. in Mathematical Statistics, and a B.S. in Psychology and Mathematics, all from the University of Illinois.

Dr. Manfred Steffen, ETS Principal Research Scientist, will assist with the psychometric needs of the SBAC program. He is currently responsible for planning and directing the psychometric and delivery components of the revised Graduate Record Examinations® (GRE®). He created and implemented the processes for the long-term management and maintenance of the revised GRE®. Dr. Steffen has also published articles and conducted presentations on a diverse range of topics, including adaptive testing and item model design. He has over 15 years of combined years working at ETS. Prior to accepting his most recent position at ETS, he was the director of analysis methods and applications at CTB/McGraw-Hill. He earned his Ph.D. in Educational Measurement and Statistics at the University of Iowa, and his M.A. in Mathematics Education and B.A. in Mathematics, both from Stetson University in DeLand, FL.

Dr. Venessa Lall, ETS Senior Psychometric Director, will contribute her psychometric and management expertise to the SBAC. She has been with ETS since 2003 and currently manages up to 40 psychometric staff who deliver research and statistical analysis activities in support of several large, complex testing programs. A sampling of these programs includes the California Alternate Performance Assessment (CAPA); California High School Exit Examination (CAHSEE); California Modified Assessment (CMA); California Standardized Testing and Reporting (STAR) program; Educational Records Bureau – Computer Adaptive Achievement Test (ERB CAAT); Maryland High School Assessment (MHSA); Standards Based Testing in Spanish (STS); Washington Comprehensive Assessment Program (WCAP); Test of English as a Foreign Language (TOEFL) and Test of English for International Communication (TOEIC). In conjunction with psychometric staff and through their activities, she affirms the technical quality of products and services, and she establishes and maintains consistent, efficient, statistical, and psychometric processes. Venessa’s other major duties include supervising, monitoring, and providing psychometric oversight and expert guidance to help effect compliance with client specifications, ETS policies, and sound measurement principles, all of which will enrich Wyoming’s scaling and equating project. She earned her Ph.D. in Educational Psychology with a specialization in quantitative methods from the University of Wisconsin–Madison, where she also obtained her M.S. in Statistics. In addition, she earned her M.B.A. from Edgewood College in Madison, Wisconsin, and her B.Sc. from the University of the West Indies in Trinidad.
Dr. Richard Schwarz, ETS Senior Psychometrician, has spent the last two years at ETS in K-12 programs serving as a senior psychometric advisor to a variety of state level programs that range from NCLB to high school graduation programs. In that capacity, he consults on technical issues both internally and externally with other psychometricians and clients on a variety of issues related to large-scale assessments. Prior to ETS, he was a consultant to various testing organizations and the Department of Defense. He spent 15 years in the capacity of a psychometrician at CTB/McGraw-Hill. Dr. Schwarz was the lead other psychometricians involved in the vertical scaling of several well-known K-12 norm-referenced tests, state testing programs, and international assessments. He has gathered extensive related to the IRT scaling and equating in many different contexts. He was also the primary architect on a system for scoring constructed-response for large-scale programs and provided consultation to many states on issues related to constructed-response in that capacity. Finally, he was a primary researcher related to automated scoring algorithms at McGraw-Hill.

Dr. Schwarz has written many papers and presentations on a variety of technical issues. Chief among his interests has been constructed-response. He is nearing completion on a book pertaining to constructed-response on issues ranging from rater agreement statistics, scaling and equating to automated scoring. He has published articles and presentations concerning constructed-response ranging from multidimensional IRT to IRT estimated reliability for polytomous items. He earned his Ph.D. in Measurement and his M.A. in Educational Psychology, both from the University of Arizona, and his B.S. in Zoology & Psychology from the University of Nebraska.

Dr. Ying Lu, ETS Senior Psychometrician, will provide psychometric support for SBAC. Since joining ETS in 2004, she has been involved in planning, designing, coordinating, and conducting research projects and operational work with the testing programs of California Standards Tests and Standards-based Tests in Spanish, which are components of the California K–12 assessment. Dr. Lu’s primary research interests are IRT, model fit, and growth. She earned her Ed.D. in Research and Evaluation Method and M.S. in Statistics at the University of Massachusetts, Amherst.

Bihua Xiang, Psychometric Project Coordinator, will oversee the program team, and internal and external consultants for SBAC. He is currently a psychometrics analyst in the K–12 Assessment Center in ETS’s Research & Development Division, working on the Washington Comprehensive Assessment Program. He is responsible for conducting statistical analysis, including test score equating, item analysis, and reliability analyses; as well as project management and large-scale testing program coordination, and psychometric review of analyses. Prior to joining ETS, he worked as a Research Assistant at the University of Alberta in support of SAT research and on-line lectures and video conferencing. Mr. Xiang’s other experience includes working as a multimedia designer and programmer for a communications company, and as senior level technical support in Canada. He also taught for three years at the Dalian University of Technology in the People’s Republic of China. Mr. Xiang earned his M.Ed. in
Media and Technology from the University of Alberta, and his B.A., graduating with honors from the Dalian University of Technology. He is a member of the National Council on Measurement in Education (NCME), and is a Microsoft Certified Professional and Certified Base Programmer for SAS 9.

**Jeannine Gilsdorf, ETS Manager, Data Analysis and Computational Research**, will oversee technical support related to the SBAC. She joined ETS in 1988, and she manages the Technology Information and Processing Services (TIPS) organization, which is accountable for the technical needs and desktop support of all members of the ETS Statistical Analysis & Psychometric Research area. She also oversees the design, development, release, and maintenance of statistical applications, solutions, and data analyses within the same area. Ms. Gilsdorf computer competencies include SAS, VBA, Visuals Basic, and JCL. She earned her B.A. in Mathematics Education from The College of New Jersey in Trenton.

**John Bonett, ETS Principal Research Systems Specialist**, will bring his knowledge of data processing design to SBAC. Since joining ETS in 2001 he has designed and implemented applications for the Statistical Analysis department related to data processing and QC, statistical report generation, and process automation. He is a certified SAS Programmer and also has active experience with SQL and db querying, VB, C#, and .NET application design. Mr. Bonett has co-authored two papers, including “Are Basic Skills Test a Useful Signal or Unnecessary Obstacle in Teacher Education: A Look at the Data” for the *American Education Research Journal*. He earned his B.A. in Mathematics and Psychology from Duquesne University in Pittsburgh.

**Dr. Michael Kane, ETS Messick Chair for Validity Research**, will provide his expertise to the SBAC. Since joining ETS in 2009 he has been responsible for maintaining a research program on topics related to the theory and practice of validation. Prior to joining ETS he was the director of research for the National Conference of Bar Examiners from September 2001 to August 2009. From 1991 to 2001, he was a professor of kinesiology in the School of Education at the University of Wisconsin–Madison, where he taught measurement theory and practice. Before his appointment at Wisconsin, Dr. Kane was a senior research scientist at ACT, where he supervised large-scale validity studies of licensure examinations. His main research interests are in the areas of validity theory and practice, generalizability theory, and standard setting. Dr. Kane has published numerous articles in refereed academic journal. He earned his Ph.D. in Education and his M.S. in Statistics, both from Stanford University, and his B. A. in Physics from Manhattan College.

**Dr. Jonathan Weeks, ETS Associate Research Scientist**, will consult on the implementation of multidimensional item response theory (MIRT) and the use of MIRT software along with growth modeling and analyses for SBAC. Prior to his arrival at ETS, Dr. Weeks worked as a data analyst, assessment director, and statistical consultant for several school districts on projects emphasizing the development and implementation of growth models. His interest in growth is complemented by a focus on methodological issues associated with test linking and vertical
Dr. Cara Cahalan Laitusis, ETS Senior Research Scientist, will be a technical advisor to ETS staff on the area of comparability and validity of test scores taken with and without test accommodations and accessibility features. Currently, Dr. Laitusis is the principal investigator and project director for three grants from the U.S. Department of Education, all of which focus on improving state assessments for students with visual impairments, blindness, learning disabilities, and mild to moderate cognitive impairments. She joined ETS in 1998 and her applied specializations are in curriculum-based assessment and the diagnosis and treatment of students with learning disabilities. She has been involved in research on the validity and fairness of assessments for all test takers. These projects included field testing of new item types for students with disabilities on both the SAT and GRE, examining the validity of testing accommodations for students with disabilities on a variety of tests, investigating gender differences in mathematical problem solving, and examining the comparability of paper- and computer-based test formats between gender and ethnic groups. Dr. Laitusis has authored numerous research articles and co-edited the book Large Scale Assessment and Accommodations: What Works?, which was published by the Council for Exceptional Children in 2007. She earned a Ph.D. in Urban School Psychology and a M.S in Educational Psychology, both from Fordham University, and a B.S. in Psychology and African Studies from Trinity College, Hartford, CT.

John W. Young is a Research Director in the Center for Validity Research at ETS, and will provide consultation on all issues to address inclusion and validity studies for English Language Learners. He presently co-directs the research initiative on assessments and products for English language learners. Prior to returning to ETS in 2006, he was a faculty member for 17 years in the Educational Statistics, Measurement, and Evaluation program at Rutgers University in New Brunswick, New Jersey. He received his Ph.D. in educational measurement from Stanford University in 1989, and in 1999, he received the Early Career Contribution Award from the American Educational Research Association’s Committee on Scholars of Color in Education for his research on academic achievement and minority students. Dr. Young has published in the following academic journals: Applied Measurement in Education, Assessment in Education,

**Dr. Rick Tannenbaum, ETS Director of Research, Center for Validity Research,** will act as consultant for standard setting activities. He joined ETS in 1988 as an associate research scientist and has since served in roles of increasing responsibility, currently as director of research in our Center for Validity Research. His research unit within the Center for Validity Research supports the content-based validity of ETS’s *Praxis™* Tests for educator certification, and it also designs and conducts standard setting studies across our testing programs. His areas of expertise include licensure testing and certification, standard setting, and validity. In addition, he has authored or co-authored more than 50 technical reports and four book chapters, and has been published in journals such as *Educational and Psychological Measurement*, *Journal of Applied Psychology*, and *Journal of Personnel Evaluation in Education*. He will bring his more than 23 years of professional and academic expertise in the educational testing field to your program. He earned his Ph.D. in Industrial/Organizational Psychology from Old Dominion University, his M.S. in Industrial Psychology from Rensselaer Polytechnic Institute, and his B.A. in Psychology from the State University of New York at Stony Brook.

**Dr. Patricia Baron, Senior Research Project Manager,** has worked with ETS since 1987, and she brings these more than 24 years of educational testing experience to the SBAC. Since 2006, she has served as the standard setting director, researcher, and lead facilitator in our Center for Validity Research. In this role, she directs standard setting for all of our K–12 testing programs for ETS, including the California Standardized Testing and Reporting (STAR) program. For the past five years, she also has focused on research in replicability and factors contributing to variability in standard setting. Significantly, she completed design and implementation of a standard setting tool for the Bookmark method, which provides a mechanism for expedited analysis and reporting with high quality assurance standards. Before transitioning into her current position, she worked as the director of Government Relations and Assessment Services, from 2005–2006. Previously, from 1989–2005, she was a senior psychometrician in our Research and Development (R&D) Division. During her 24 years at ETS, she has been the lead psychometrician on high-stakes undergraduate and graduate admissions tests, outcome assessments for college and higher-level programs, and a national assessment for Qatar. She has led development of the vertical scale and test design and helped plan standard setting for Qatar in Arabic and English. From 1992–1995, she served as assistant editor of *Educational Measurement: Issues and Practice*. She earned her Ed.D. in educational psychology, with a specialization in educational statistics and measurement, from Rutgers University, where she also earned her B.A. in psychology.

**Dr. E. Caroline Wylie, ETS Managing Research Scientist,** will provide her expertise for SBAC. She joined ETS in 1997 and works within the Cognition and Learning Sciences group. Currently, she is the co-principal investigator focusing on the validation of a set of formative assessments in mathematics that are built around learning progressions. The author of several published
articles within the field, Dr. Wyle’s research centers on issues of teaching quality and the use of formative assessment as a mechanism for improving teaching and learning in classrooms. She earned her Ph.D. in Educational Assessment from Queen’s University, Belfast, and her postgraduate certificate in Teaching (mathematics and information technology). She earned her B.Sci. in Applied Mathematics and Physics, also from Queen’s University.

Dr. Derrick Higgins, ETS Research Director, will bring his knowledge to the SBAC. He is responsible for development and oversight of technologies related to the automated scoring of constructed-response test items, student learning tools, and tools for automated development of assessment materials. He has worked for Microsoft as a natural language processing research intern and for Inxight Software as a computational linguist. He was also a linguistic consultant for the American and French Research on the Treasury of the French Language (ARTFL) Project in 1999 and for Argonne National Laboratories in 2002. Dr. Higgins has worked in ETS’ Research division for over eight years, first in a scientist role (contributing to ETS’ e-rater and SpeechRater scoring systems, as well as tools for automated item generation), and then assuming leadership responsibility for the Natural Language Processing and Speech group. He has published several papers in linguistics, computational linguistics and assessment, and has presented the work of ETS’ NLP & Speech group widely at policy and research forums. Dr. Higgins earned his Ph.D. and B.A. in linguistics from the University of Chicago and has also studied at the University of Michigan, Humboldt University of Berlin, and Free University of Berlin.

Dr. F. Jay Breyer, ETS Principal Research Project Manager, will oversee project work for SBAC. He joined ETS Research in 2010 is responsible for psychometric evaluation and work on the use of automated scoring systems and models. He has led the development of an internal bank of items for use with c-rater™ and m-rater™ for low-stakes formative assessment, as well as for use with SpeechRater in EFL Teacher Certification and TOEFL® Junior. He also manages the documentation and work processes and procedures related to the use of e-rater® on various ETS examination programs, including Graduate Record Examinations® (GRE®), GRE® ScoreItNow, Praxis™, and Test of English as a Foreign Language™ (TOEFL®). He has several articles published and has conducted numerous presentations on a range of psychometric issues. Prior to coming to ETS in 1988, Dr. Breyer was the executive director of Psychometric Consulting Services. He earned his Ph.D. in Psychometrics from Fordham University, his M.A. in Psychology from St. John’s University, and his B.A. in Psychology from Cathedral University.

Dr. Chaitanya Ramineni, ETS Associate Research Scientist, will assist with research for the SBAC. Since joining ETS in 2009, her research has included an examination of the automated scoring of constructed response items and analyses of the Graduate Record Examinations®, the Test of English as a Foreign Language™ (TOEFL®) and e-rater®. Prior to joining ETS, she was a measurement intern at the National Board of Medical Examiners (NBME) in Philadelphia. In 2009 Dr. Ramineni won the Thomas Hale Ham New Investigator Award for research in medical education from the Association of American Medical Colleges. She earned her Ph.D. in Education (Research, Methodology, and Evaluation) and her M.S. in Computer & Information Sciences,
both from the University of Delaware, and her B.E. in Computer Science & Engineering from Rajiv Gandhi University, Bhopal, India.

**Consultants**

Dr. Wendy Yen, ETS External Consultant, will provide psychometric support for SBAC. Before retiring from ETS as a Distinguished Presidential Appointee in 2011, she provided senior oversight of K–12 assessment, and consulted with educators on the design and implementation of high quality assessment programs. Prior to joining ETS in 2000, she was vice president of research at CTB/McGraw-Hill, where she oversaw the technical quality of all their published assessments, including the widely used TerraNova, SUPERA, and California Achievement Tests (CAT) achievement batteries. She has provided technical guidance for more than 30 customized state assessment programs and consulted on the National Assessment of Educational Progress (NAEP). She received the 2008 American Educational Research Association/American College Testing E.F. Lindquist Award for lifetime contributions to testing and measurement. In addition, she has served as president of the National Council on Measurement in Education and as trustee of the Psychometric Society, and on committees for the National Academy of Sciences and the National Assessment Governing Board. She also served as editor of the *Journal of Educational Measurement*, and as associate editor of *Psychometrika*. She is the coauthor of *Introduction to Measurement Theory*, which has been in print for more than 25 years, and author of more than 100 professional presentations and publications, primarily in the areas of test scaling and equating. She earned her Ph.D. in Mathematical Psychology, her M.A. in Applied Statistics, and her A.B. in Psychology, all from the University of California at Berkeley.

Dr. Daniel F. McCaffrey, ETS External Consultant, will provide psychometric support for SBAC. He currently is a senior psychometrician at the RAND Corporation, where he holds the PNC Chair in Policy Analysis. His research topics have included score estimation, educational evaluation, and teacher effectiveness. Prior to joining the RAND Corporation he was a professor in the Department of Statistics at the University of Pittsburgh. Dr. McCaffrey has published widely, including articles in the *Journal of Educational and Behavioral Statistics* and the *Journal of Research on Educational Effectiveness*. He earned his Ph.D. in Statistics from North Carolina University and his B.A. in Mathematics and Economics from Mount Saint Mary’s College.

**Measured Progress Staff Qualifications and Experience**

Our staffing table for the SBAC-05 contract show all named staff from Measured Progress, their primary responsibility (role) for this contract, and the amount of time each will be assigned to this project for the specified contract period.

We confirm that the staff members we have named in our proposal will perform the assigned work, and that we are to obtain prior approval from SBAC Leadership before carrying out any subsequent staff substitution.
**Measured Progress Staff Biographies**

**Dr. Michael Nering, Measured Progress Assistant Vice President of the Research and Analysis**, is responsible for overseeing the daily operations of all psychometric functions. Drawing upon years of psychometric experience, Dr. Nering provides psychometric support and initiates process improvement. Dr. Nering also stays current within the psychometric community by presenting at industry conferences and publishing various articles and books.

Dr. Nering’s research interests include person fit, item response theory, computer-based testing, and equating. He has presented and published numerous articles on a wide range of psychometric topics, and he is actively involved in the research community in various capacities.

Dr. Nering is a member of the National Council of Measurement in Education, American Educational Research Association, American Psychological Association, and the Psychometric Society. For the AERA 2005 conference he was a program chair for Division D – Measurement and Research Methodology. He has also served as treasurer of the Psychometric Society. In addition, Dr. Nering has served as reviewer for several peer journals, including the Journal of Educational Measurement, Applied Psychological Measurement, Psychometrika, and the Journal of Experimental Education.

Dr. Nering has a Ph.D. in Psychology with a specialization in Psychometric Methods from the University of Minnesota and a Bachelor's degree in Psychology from Kent State University, Kent, Ohio.

**Dr. Jennifer Dunn, Measured Progress Assistant Director of the Psychometrics and Research division**, is responsible for overseeing the daily operations of all psychometric functions. Drawing upon years of psychometric experience, Dr. Dunn provides psychometric support and initiates process improvements. Dr. Dunn also stays current within the psychometric community by presenting at industry conferences and publishing various articles.

Prior to joining Measured Progress Dr. Dunn worked at The National Center for the Improvement of Educational Assessment for three years. Since joining Measured Progress, Dr. Dunn has been involved in overseeing and implementing a variety of complex psychometric activities, most notably standard settings and technical reports. She has frequently represented measured progress at technical advisory meetings and has consulted on a variety of special projects both internally and externally. Dr. Dunn’s evolving responsibilities have led to frequent interactions and close working relationships with clients and representatives from Measured Progress’ functional groups.

In addition to her experience in psychometrics, Dr. Dunn has taught at the college level, most recently, Psychometrics II at Boston College.
Dr. Dunn has a Ph.D. in Measurement and Evaluation from the University of Toronto, Ontario, Canada; an M.A. in Measurement and Evaluation from the University of Toronto, Ontario, Canada; and B.S. degrees in Kinesiology and Psychology from the University of Waterloo, Ontario, Canada.

Below are additional biographical summaries for consultants from Measured Progress that we are proposing to complement the project staff, adding critical niche areas and layers of expertise that will further enhance the solution we offer SBAC.

**Dr. Michael Russell, Measured Progress Nimble Innovation Lab Vice President**, manages and leads research and development projects that focus on enhancing the validity of educational assessments through the use of digital technologies. He is trained in educational measurement and has eighteen years of experience conducting leading edge research and development of innovations to assessment. He has developed strong leadership, communication, and project management skills that support collaborative efforts to improve the quality of educational assessments. This work has resulted in more than 50 scholarly articles and five books on educational assessment. Areas of specialty for Dr. Russell include computer-based testing, accessibility, technology-enhanced assessment, and interoperability standards. He earned his Ph.D. in Educational Research from Boston College, his M.S. in Secondary Education from Boston College, and his B.A. in History from Brown University.

**Dr. Luz Bay, Measured Progress Assistant Vice President of Client Services**, leads the team at Measured Progress to assist the National Assessment Governing Board (NAGB) in setting achievement levels for the National Assessment of Educational Progress (NAEP) in Writing for grades 8 and 12 and conducting a judgmental standard setting study on grade 12 academic preparedness in reading and mathematics. For these two projects, Dr. Bay provided vision and guidance in the development of cutting edge technology for computer-based implementation of popular standard setting methods. The Computer-Aided Bookmarking (CAB) was successfully launched in the spring of 2011 and the Body-of-Work Technological Integration and Enhancements (BoWTIE) will be used in November, 2011.

Most recently, Dr. Bay led the Data Services and Score Reporting department where she oversaw all the work related to data processing, data analysis, report design, and data quality assurance of assessment results, as well as report design and the production of technical reports. During her tenure in this position, Measured Progress has gained a reputation for data integrity and reporting accurate results. In this role she also provided psychometric expertise and support for the application of quantitative methodologies to state assessment contracts and developed methods to conduct quality assurance. Her career at Measured Progress and extensive history of leadership, having moved quickly from Senior Psychometrician, Manager of Data Analysis, Assistant Director of Measurement, Design, and Analysis, and Director of Data Processing and Analysis.
Prior to joining Measured Progress, Dr. Bay was a psychometrician in the Research Division at ACT, Inc. for five years. She served as the assistant project director for the National Assessment of Educational Progress (NAEP) Achievement Levels-Setting (ALS) Project for the National Assessment Governing Board (NAGB). She was involved in both the operational and research aspects of setting achievement levels for the 1994 NAEP in geography and U.S. History; the 1996 NAEP in science; and the 1998 NAEP in civics and writing, which included an extensive research agenda that was her primary responsibility. In addition to her involvement in the research and operational aspects of the project, she has also been a facilitator during the achievement levels-setting meetings and in charge of training other facilitators. She has also written numerous reports presented to the NAEP Technical Advisory Committee on Standard Setting (TACSS) on results of research studies related to the NAEP standard setting process. In her previous position, she played a major role in conceptualizing, designing, and implementing research studies for the NAEP ALS project.

Dr. Bay earned her Ph.D. in Educational Measurement and Statistics and M.S. in Mathematics, both from Southern Illinois University, and her B.S. in Mathematics from the University of the Philippines.