Supporting Statement A

# The Social Security Administration (SSA)- National Institutes of Health (NIH) Collaboration to Improve the Disability Determination Process: Calibration II & Predictive Validity Testing of Item Response Theory – Computer Adaptive Testing Tools (IRT-CAT)

NIH/CC/RMD

Date 08/11/2014


| | |
|---|---|
| Name: | Daniel Hobbs<br>Management Analyst<br>Epidemiology & Biostatistics Section<br>Rehabilitation Medicine, CC, NIH |
| Address: | 10 Center Drive, Bld. 10, 1-1469 |
| Telephone: | 301.496.3817 |
| Fax: | 301.480.0415 |
| Email: | Daniel.Hobbs@nih.gov |

## TABLE OF CONTENTS

## A. JUSTIFICATION

**LIST OF ATTACHMENTS**:

**Attachment 1 – Authorizing Legislation**

**Attachment 2 – Pre-notification letter**

**Attachment 3 – Claimant IRB-Approved Consent**

**Attachment 4 – Complete Item Bank for Daily Activities & Learning and Applying Knowledge Functional Assessment Batteries (Survey 1)**

**Attachment 5– Short-form versions of Physical Function & Behavioral Health Functional Assessment Batteries including Replenishment Items (Survey 2)**

**Attachment 6 – Normative Survey 1 (including consent)**

**Attachment 7 – Normative Survey 2 (including consent)**

**Attachment 8 – Claimant Screener Script Survey 1**

**Attachment 9 – Claimant Screener Script Survey 2**

**Attachment 10 – Westat Media Sanitization & Disposal policy**

**Attachment 11 – Data Security- Authorization to Operate**

**Attachment 12 – NIH Privacy Impact Assessment (PIA)**

**Attachment 13– Institutional Review Board Approval Notification**

**Attachment 14 – NIH Institutional Approval (Executed Reliance Agreement)**

## A.1 CIRCUMSTANCES MAKING THE COLLECTION OF INFORMATION NECESSARY

The National Institutes of Health, Clinical Center, Rehabilitation Medicine Department seeks OMB clearance to conduct calibration, predictive validity and validation studies on Computer Adaptive Tests currently under development under an Interagency Agreement with the Social Security Administration to improve the Disability Insurance Program. We are requesting a 3-year clearance to conduct these activities.

The Social Security Administration (SSA) entered into an interagency agreement (IAA) with the National Institutes of Health (NIH), Clinical Research Center, Rehabilitation Medicine Department (RMD) to explore innovative methods of functional assessment to improve SSA's disability determination process.

As part of its study, NIH recommended item response theory (IRT) coupled with computer adaptive testing (CAT) as a promising approach to efficiently and consistently capture claimant functional information to assist SSA adjudicators. IRT is a framework for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, aptitudes, and other variables. It is often the preferred method for the development of tests such as the Graduate Record Examination (GRE) and the Graduate Management Admission Test (GMAT).

Likewise, Computerized Adaptive Testing (CAT) is a form of computer-based testing that tailors question selection based upon the examinee's ability level. A CAT is programmed to initially present a question from the mid-range of a hierarchically ordered list of questions and then select subsequent questions at an appropriate level based on the examinee's previous answers. In contrast to traditional, fixed form functional tests that ask the same questions of everyone regardless of how the respondent answers; CAT instruments, like a skilled clinician, tailor the assessment by asking only the most informative questions based on a person's response to previous questions. Thus, fewer questions (in total) are needed since the questions are selected based on the individual's level of function and test results can be computed in a matter of minutes with just a few questions.

The Epidemiology and Biostatistics section in RMD will be collecting information through a contractor (Boston University- Health and Disability Research Institute (BU-HDRI)) and subcontractor (Westat, Inc.) for calibration and determination of predictive validity of the CAT tools under development to assist in the SSA disability determination process. The utilization of CAT technology could potentially allow the SSA to collect

more relevant and precise data about human functioning in a faster, more efficient fashion.

The NIH/RMD awarded an initial contract to the Boston University Health and Disability Research Institute (BU-HDRI) in 2009 to evaluate the feasibility of integrating this promising new testing method into the SSA's data collection processes. In order to understand distinct factors influencing work, individual capabilities as well as workplace demands and critical features of the workplace environment must be captured. The contract with Boston University encompasses CAT development to capture the —person side of this interaction, in other words, the assessment of individual capabilities.

The development of CAT tools is a sequentially dependent process. Therefore, each step of CAT tool development proceeds in an ordered fashion; one step must be completed before advancing to the next step. The first step of the process is item pool development. This step encompasses working with content experts, examining literature and reviewing other models/taxonomies to develop item pool content and structure. The next step is to calibrate the items of each pool. Statistical analyses are conducted on data collected from samples of persons similar to the intended audience for the instrument. The objective is to assess the psychometric properties of the items in the pool. The final step of developing CAT tools is to validate and bookmark the instrument, necessary to demonstrate defensibility and to denote cut-points that may aid in disability evaluation decision-making. While the initial contract will develop multiple CAT tool instruments, the content of each instrument is unique and development of each instrument must follow the sequential process.

Initially BU focused on the identification of functional domains appropriate for CAT instrument development, relevant to SSA's need to determine work disability. *Physical Function* and *Behavioral Health* were selected as the initial domains for development. The selection was motivated by prior work on mobility CAT instruments that could be tailored for SSA's needs; and, by SSA's desire for improved approaches to evaluate claimants with cognitive and mental health conditions. The remaining functional domains will be assessed by two CAT instruments to assess *Learning and Applying Knowledge* which includes aspects of cognition, communication, language and social interactions as well as *Daily Activities* including aspects of self care, social appropriateness, independent living skills and transportation.

These CAT instruments, being developed in each of the four domains in an effort to improve the SSA disability determination process, comprise the complete catalogue of Functional Assessment Batteries (FABs).

Development of the item pools is an iterative process influenced by the literature, existing instruments, content experts, focus groups, and cognitive testing. BU developed detailed schematics of the content models for both domains and operationally defined terminology to facilitate clarity and enhance precision of the sub-domains encompassed within each model.

Disability, in this circumstance—SSA's perspective of work disability, is the interaction between the functioning of the whole person and environmental demand. The assessment of functioning provides SSA a mechanism to integrate contemporary perspectives of disablement into disability program processes. The use of IRT/CAT assessments may allow SSA to capture functional information in a more precise, efficient and comprehensive manner. This may improve the uniformity of decision-making and potentially reduce program costs by informing decision-making earlier in the evaluation process.

Calibration and Predictive Validity

A calibration study is a field study of item content and structure conducted with samples of respondents representing the intended users of the CAT instruments. Item pool development and subsequent item calibration are unique for each CAT tool and for the target population for which they are developed. Sample size is determined based on the statistical need to support a series of confirmatory factor analyses and to perform statistical modeling. Inadequate sample size may lead to inaccurate and unstable statistical outcomes.

The SSA-NIH-BU team collaborated with the subcontracted national survey firm, Westat, to establish the calibration design and sampling strategy. SSA's Office of Data Analysis will extract from SSA automated records claims submitted within the last two months, and provide from existing SSA administrative data.

From this dataset, Westat will assign a geographic variable to the data to classify claimants into urban or rural categories. The sample will then be stratified by urban/rural status across the 10 national SSA office regions. A randomly selected subsample of 20,000 will then be drawn.

Westat will contact and notify the claimants in about the study with a pre-notification letter **(Attachment 2)** and consent form **(Attachment 3)** for interested participants. Consent will be obtained at the beginning of each survey administration.

Participant data will be retained until a disability determination is made or until 2 years following initial study enrollment. CAT scores will be examined relative to determination

outcomes controlling for other potentially influential factors. This will permit the research team to ascertain the relative predictive validity of the instruments.

In addition to collecting data from SSA claimants, data from a normative sample will be collected. The normative sample data allows the research team to expand the breadth of each scale developed compared to use of claimant data alone. This will reduce ceiling effects and broaden the overall applicability of each CAT scale developed. Secondly, having calibration data from a normative sample of adults in the country provides a useful reference population against which SSA claimants can be compared. This allows SSA to better characterize their population of claimants over time.

The normative national sample will be obtained using sample matching; a methodology pioneered by YouGov Polimetrix, Inc. (YGP; Palo Alto, CA) whereby samples representative of a study-appropriate target population can be constructed from large but unrepresentative pools of opt-in survey respondents. The enumeration of the target population would in traditional sampling be known as the sampling frame and would serve as the source from which the sample would be drawn. This is not the case in sample matching, which instead proceeds in two-stages. First, a random sample is drawn from the enumeration of the target population. A simple random sample (SRS) could be drawn; but in practice, the efficiency of the procedure can be improved by using stratified sampling. YGP typically stratifies on race (utilizing OMB standardized categories), gender and age, and then draws a SRS from each of the mutually exclusive and exhaustive groups formed by the simultaneous cross-classification of the population on these three attributes. The SRS from each category is combined to form the stratified target sample. If the number of respondents selected in each stratum is proportional to their frequency in the target population, then the sample is self-representing.

Conventionally, one would then attempt to contact the respondents in the target sample. However, there is no economical way of reaching most members of the target sample, as they have not provided their email addresses and many do not have a listed phone number, and those that do, may not agree to be interviewed. Instead, for each member of the target sample, YGP will select one or more matching members from their pool of opt-in respondents. This pool has been recruited by a variety of means and currently numbers approximately 1.5 million. Of course, data drawn from this pool would not be representative of any particular population; individuals who opt-in for taking web surveys have different demographics than either the population of all internet users or the population of all adults. Rather, the matching methodology is required to produce usable samples for individual studies. Matching is done on a large set of variables available in both the population enumeration database and the opt-in panel. The purpose of the matching is to find an available respondent who is as similar as possible to the selected member of the target sample. YGP employs a proximity matching method whereby a

distance function is computed for each attribute to define the degree of "closeness" between each individual in the target sample (x) and those in the opt-in survey panel (y). Typically, the distance function is the simple absolute value of the difference, |x-y|, and the overall distance between a member of the target sample and a member of the panel is a sum of the distance functions for each attribute being used in the matching. The distance functions can be weighted and then summed if particular variables are thought to be more important for a given study. For this study, matching was done on gender, racial/ethnic background, age, education and employment status, weighted equally.

YGP adjusts for anticipated non-response by selecting multiple best matches in the opt-in panel for each member of the target sample. The number of matches is determined by using a hazard model to estimate the probability that an opt-in panelist will respond by the end of the data collection period, and increasing the number of panelists matched to the member of the target sample until that response probability is >=1. Although internet use was initially concentrated in the more affluent and better-educated segments of the population, this "digital divide" has been substantially reduced such that according to the United States Census Bureau, over three-quarters of the adult population now have access to the internet either at home, work or school.

YGP will use web survey administration to deliver the same item pool surveys used with SSA claimants. Participants will be consented before beginning the online survey.

Validation Phase

To validate the new instruments, they must be tested against gold-standard legacy instruments to determine their effectiveness in capturing functional ability. For this phase of the study, an opt-in sample of 500 self-classified "work disabled" adults, recruited by the survey firm YouGov Polimetrix will complete eight (8) functional assessment legacy instruments: VR-36, PROMIS Physical Function SF 10 item, PROMIS Applied Cognition: Abilities and General Concerns short forms, BASIS 24 (omitting drug and alcohol questions), Generalized Self Efficacy Scale (GSE), Functional Independence Measure (FIR-SR) Self Report, LaTrobe Communication Questionnaire, and AM-PAC.

The sample will then complete the four developed CAT instruments (Physical Functioning, Behavioral Health, Learning & Applying Knowledge, Activities of Daily Living) during the same contact. It is estimated, based on data from previous studies, that the participants will take no more than one hour to complete the legacy instruments and 30 minutes to complete the four CAT instruments.

Reliability Testing

The reliability portion of this study will examine the outcome consistency of the developed CAT instruments, and aims to identify and resolve potential CAT reliability issues.

Specific study primary objectives for this phase include:

1. To examine the existing CAT instruments' consistency and reliability in scoring function of individuals in a normative population sample and those with disabilities

2. To differentiate between the fluctuation of individual CAT scores as functional status changes occur in claimants over a period of time and limitations of the CAT instruments

YouGov will administer two assessments of *Learning & Applying Knowledge* and *Daily Activities* to a normative study population of 400 US adults and 400 self reported permanently disabled US adults for the initial contact. 300 individuals in each of these two samples will be contacted a second time, for a second administration of the same instruments. We anticipate, based on findings of previous validity testing for earlier CAT instruments, a 25% attrition rate. Respondents will also be administered the legacy assessment VR12 at each time point. All assessments will be administered electronically through the company's website. The study participants will be an opt-in pool of voluntary respondents, recruited through the standard practices of private survey firms, including incentivizing through a point-based system.

The two instruments under development, along with the VR12 assessment will be administered to participants twice. The second time point will be administered seven (7) days following the initial administration. Administration will occur on a rolling basis, to ensure that all study participants are tested at a uniform interval. This second series of tests will allow for analysis of any changes in scoring, specifically assessing the ability of the instruments to consistently score an individual's function.

To ascertain those individuals whose functional status changes between administrations of the two tests, the legacy instrument VR12, considered the current 'gold- standard' in functional testing, will be administered to all survey respondents. This will allow researchers to account for expected variations in CAT scores among those whose functional status has changed between the administrations of the two rounds, differentiating these scores from variations in scores due to limitations of the CAT instruments.

<u>Replenishment Phase</u>

As a final phase of development, CAT instruments must undergo replenishment to refine their scope and address any disparities in the instruments' ability to classify function across the entire range of human functioning. For this activity, replenishment items developed by content experts will first be cognitively tested. It is anticipated 20 individuals comparable to the work-disabled target population of SSA claimants, recruited through the Boston University Health & Disability Research Institute, will complete an interview to elicit problems with item interpretation, wording and ambiguity in response options for questions included in the Learning & Applying Knowledge (LAK) and Activities of Daily Living (ADL) CAT instruments.
After completion of psychometric testing on the replenishment items, 1,000 SSA claimants for disability benefits will complete the short-form versions of the two new instruments and replenishment items.

This initiative is authorized by section 1110(a) of the Social Security Act, as amended (42 U.S.C. § 1310(a)). **(Attachment 1)**

NIH CRC/RMD entered into the IAA with SSA under the authority of 42 U.S.C. §§ 241, 282, and 284.

## A.2 PURPOSE AND USE OF THE INFORMATION COLLECTION

Data for the study described will be collected by the NIH Clinical Center through a contract with Boston University and sub-contracts with Westat and YouGovPolimetrix (Polimetrix), a survey research firm based in Palo Alto, CA. This information will be used to develop the BU-HDR CAT instruments. The proposed information collection will support tool development and psychometric testing. The calibration/predictive validity phase is a longitudinal, observational study. For the initial survey, the *Daily Activities* and *Learning and Applying Knowledge* FAB item pools/ Survey 1 will be administered **(Attachment 4)**. During the follow-up survey, the short form versions of the *Physical Function and Behavioral Health* FABs/ Survey 2 will be administered, allowing for testing and scaling of replenishment items **(Attachment 5)**.

Participant data will be retained until a determination is made or until 2 years following initial study enrollment. FAB scores will be examined relative to determination outcomes controlling for other potentially influential factors. This will permit the research team to ascertain the relative predictive validity of the FAB instruments.

Purpose of the Calibration II/Predictive Validity Phase

1. Confirm which questions are part of a particular content domain
2. Determine which questions are clear, concise and discriminate among factors in each construct and should be included in the FAB
3. Identify properties that are needed for FAB administration
4. Verify that all FAB model assumptions are met
5. Determine the extent to which FAB scores in the physical function, behavioral health, daily activity, and learning and applying knowledge domains, in conjunction with other information about claimants, predict subsequent SSA disability determination outcomes (the criterion measure)

Specific objectives

1. In conjunction with other information collected by SSA about claimants, determine whether FAB scores can augment the accuracy of actual SSA disability determinations;
2. Identify how the use of FAB scores, collected early in the application process, can augment adjudication of actual SSA determinations;
3. Determine which functional domains are most related to disability determination outcomes.
    a. Explore relationships between FAB outcomes relative to Listings outcomes

<u>Objectives of the Validation Phase</u>

1. Test the instruments against gold-standard legacy instruments to determine their effectiveness in capturing functional ability

<u>Objectives of Reliability Testing</u>

1. To examine the existing CAT instruments' consistency and reliability in scoring function of individuals in a normative population sample and those with disabilities

2. To differentiate between the fluctuation of individual CAT scores as functional status changes occur in claimants over a period of time and limitations of the CAT instruments

<u>Objectives of Replenishment Phase</u>

1. Refine the scope of the instruments' item banks and address any disparities in the instruments' ability to classify function across the entire range of human functioning.

Calibration/ Predictive Validity Study Component

NIH proposes the following process for information collection relative to the Calibration phase, that will inform study design and procedure development for the subsequent phases of CAT development:

The SSA-NIH-BU/Westat team collaborated to establish the calibration design and sampling strategy. SSA's Office of Data Analysis will extract from SSA automated records claims submitted within the last two months and provide basic claimant contact information from existing SSA administrative and programmatic datasets. From this dataset, Westat will assign a geographic variable to the data to classify claimants into urban or rural categories. The sample will then be cut and stratified by urban/rural status across the 10 national SSA office regions. This will result in an initial claimant pool of 20,000.

Westat will mail a pre-notification package to the randomly selected sample of claimants. A Westat interviewer will contact each interested claimant by telephone to confirm eligibility by use of a screener **(Attachment 8)** and their willingness to participate in the study. The claimant will give verbal consent if he/she agrees to participate. If the claimant completes the web survey, there is a check box provided that the claimant must click to indicate consent in order to begin answering the survey questions. A Westat interviewer will contact each of the original study participants a second time, and confirm eligibility with a second screener **(Attachment 9).**

We estimate 3,500 claimants will complete the first survey. During the first contact, a claimant will respond to the 2 new item banks (1) *Daily Activities* and (2) *Learning and Applying Knowledge* **(Attachment 4)**. We estimate this will take 60 minutes, and we will provide a $20.00 incentive for voluntary participation in this study component. These individuals will then be re-contacted, approximately 1-2 weeks after the initial response, and will be administered the short form versions of the *Physical Function* and *Behavioral Health* FABs **(Attachment 5)**, as well as replenishment items for the same two domains. We estimate this second contact will be conducted with 3,000 claimants (accounting for attrition) and will take no more than 60 minutes to complete. We will provide an additional $30.00 incentive for voluntary participation in this component of the study.

The web survey system that Westat has developed for this project will not contain or be linked to claimant identifying information. Only Westat's unique identification number for each claimant along with his or her survey responses will be stored on Westat's system and the BU-CAT-SMS. No personally identifiable information (PII) will be collected from individual respondents. Interviewers who complete the survey over the

phone with claimants will access the web survey in the very same way the claimants would if the claimant were completing the questionnaire online. The interviewers will navigate from their calling screens to a web browser, access the web survey with the URL, and then, enter in the claimants' unique access codes, and record the claimants' responses to the web survey. When the interviewer accesses the web survey, he or she will read aloud over the phone to the claimant all introductions and instructions. The interviewer will enter the respondent's answers directly into the web survey.

In addition to collecting data from SSA claimants, data from a normative sample will be collected. The normative sample data allow the research team to expand the breadth of each scale developed compared to use of claimant data alone. This will reduce ceiling effects and broaden the overall applicability of each CAT scale developed. Secondly, having calibration data from a normative sample of adults in the country provides a useful reference population against which SSA claimants can be compared. This allows SSA to better characterize their population of claimants over time.

The normative national sample of 2,000 will be obtained using sample matching described previously in this application. They will be administered (via YouGov website) the same two surveys as the claimant population **(Attachments 6 & 7).** Participants will be consented before beginning the survey, in the same contact. It is estimated these surveys, considering mode of administration and industry averages for similar length surveys, will require no more than 45 minutes each to complete.

Predictive Validity Longitudinal Study Component

Following completion of calibration study data analysis, the complete de-identified dataset will be sent on an encrypted CD via courier from BU-HDR to NIH. Upon receipt by NIH, an approved project analyst with appropriate SSA clearance will contact BU for the password and download the data onto the NIH secure network. The CD will then be shredded following industry standard media sanitation practices **(Attachment 10)**. The calibration data for the SSA claimant participants will then be linked to their SSA electronic folder number using the unique study ID originally assigned by Westat. NIH will be the only entity with access to both SSA electronic folder number and Westat-assigned study ID for each study participant.

Claims status follow-up with SSA will be performed by the NIH scientific team, **without additional claimant contact**, up to 2 years after the second calibration study contact. Determination outcomes will be extracted from the electronic folders of the entire sample (13,260) received from SSA using their electronic folder record numbers. This will ensure SSA does not know who ultimately participates in the study, as guaranteed in the

consent document. The variables to be extracted in this follow-up will include final determination status, if available. Study participants will incur no additional burden during this phase of the study.

## A.3 USE OF INFORMATION TECHNOLOGY AND BURDEN REDUCTION

The proposed calibration study will collect all data electronically. For calibration, individuals will be able to enter their responses electronically using a link provided in the mailing sent from Westat, or call a Westat interviewer who completes the survey over the phone with claimants and will access the web survey in the very same way the claimants would if the claimant were completing the questionnaire online. The interviewers will navigate from their calling screens to a web browser, access the web survey with the URL, and then, enter in the claimants' unique access codes, and record the claimants' responses to the web survey. When the interviewer accesses the web survey, he or she will read aloud over the phone to the claimant all introductions and instructions. The interviewer will enter the respondent's answers directly into the web survey.

For the normative sample recruited by YGP, all individuals will submit responses electronically. Internet administration will reduce burden for this sample.

All of the YGP panelists have provided their e-mail so that they may receive electronic invitations to participate in surveys. Additionally, with each survey invitation they are reminded of the YGP policy on privacy, the opportunity to immediately opt-out, and of the voluntary nature of each request regardless of the survey sponsor.

Electronic data capture using computer adaptive testing technology is more efficient compared to fixed form assessment instruments and substantially reduces respondent burden.

In compliance with the NIH-SSA IAA, NIH and its contractors currently have extensive security and privacy agreements in place. The NIH information collection system that will house data for the predictive validity study has received an Authorization To Operate (ATO) **(Attachment 11).** The security planning, certification and authorization process for this project fully complies with the National Institutes of Standards and Technology (NIST) guidelines and with standards and practices outlined by the Federal Information Security Management Act of 2002 (FISMA).

## A.4 EFFORTS TO IDENTIFY DUPLICATION AND USE OF SIMILAR INFORMATION

There is no duplication of effort or similar information available for use. These are new CAT tools developed specifically for the SSA disability programs, and require calibration

prior to pilot testing. The data will be unique to the instrument and will feed back into the psychometric evaluation of the assessment instrument. Data are necessary to provide a basis and a context for structuring the items within the instrument and for assessing the predictive validity of the instruments in relation to SSA business needs.

## A.5 Impact on Small Businesses or Other Small Entities

This research will not impact Small Businesses or Other Small Entities, only individuals will be approached as potential participants.

## A.6 Consequences of Collecting the Information Less Frequently

Information will be collected from each participant as described specifically for this study. No additional contacts will occur. Combining calibration with predictive validity testing minimizes participant contact and maximizes efficiency of study resources.

## A.7 Special Circumstances Relating to the Guidelines of 5 CFR 1320.5

This project fully complies with 5 CFR 1320.5

The data collection in this project will support the development of a measurement instrument (i.e., methodological development) to assess functioning with respect to work disability. Results will not be generalized to other populations.

## A.8 Comments in Response to the Federal Register Notice and Efforts to Consult Outside Agency

**A8A** The 60-day Federal Register notice was published on April 22, 2014: Federal Register Vol. 79, No. 77. No public comments were received.

**A8B** Dr. Alan Jette, Director of the Boston University Health & Disability Research Institute has led development of these instruments. He is regarded internationally as an expert on the development of Computer Adaptive Tests utilizing Item Response Theory. He can be reached at ajette@bu.edu.

As BU began CAT development for the domains, they evaluated existing conceptual frameworks, such as the World Health Organization (WHO) International Classification of Function (ICF), in order to develop the structure for each domain. Existing conceptual frameworks were consulted including those for PROMIS, Neuro-QOL, and the NIH Toolbox (assessment of neurological and behavioral function) projects. While these existing frameworks were developed for specific populations and a different purpose, they were critical in informing domain structure for the BU CATs.

Consultation and coordination has been sought throughout the BU-HDR CAT development process. CAT tool developers examined items from existing NIH tools (such as those developed for PROMIS and Neuro-QOL) for potential inclusion in the computer adaptive tests being developed by BU. PROMIS aims to use computer adaptive testing methodology to develop ways to measure patient-reported symptoms, such as pain and fatigue, and aspects of health-related quality of life across a wide variety of chronic diseases and conditions. The Neuro-QOL is a 5-year, multi-site project funded by the National Institute of Neurological Disorders and Stroke (NINDS), which is intended to develop assessments that address dimensions of health-related quality of life that are universal to adults and children with chronic neurological disorders. Neuro-QOL is also based on patient-reported outcomes and uses CAT methods to assess pain, fatigue, emotional distress, physical function, and social function. Since both PROMIS and Neuro-QOL use CAT methods to assess patient-reported outcomes, items from these assessments could be selected for inclusion in the item banks being developed by BU.

Consultation was also sought by BU CAT development experts on writing new items to assess aspects of functioning not captured.

## A.9 EXPLANATION OF ANY PAYMENT OF GIFT TO RESPONDENTS

1. Calibration

During the first contact, a claimant will respond to the 2 new item banks (1) *Daily Activities* and (2) *Learning and Applying Knowledge/* Sample 1. We estimate this will take no more than 1 hour, and we will provide $20.00 for their time associated with voluntary participation in this study component. These individuals will then be re-contacted, approximately 1-2 weeks after the initial response, and will be administered the short-form *Physical Function* and *Behavioral Health* FABs/ Sample 2 that include replenishment items for these two domains. We estimate this will take no more than 60 minutes. We will provide an additional $30.00 for voluntary participation in this component of the study.

2. Normative Sample

YPG's goal is to provide a small thank you for a respondent's time, but not an incentive that might make survey response a financial transaction. The average survey incentive of 500 points cashes out at 50 cents, although it's not redeemable until respondents reach certain thresholds. Respondents who complete the instruments required for this normative population portion of the calibration study will earn approximately 3,000 points. For administrations of the instruments during the validation phase, reliability testing, and during replenishment, we anticipate participants will receive 500 points per

15 minutes of survey administration time. This nominal "thank you" is industry standard for completion of opt-in online survey completion through organizations such as YouGov.

## A.10 Assurance of Confidentiality Provided to Respondents

Responses to questions will remain secure to the fullest extent permitted by law. Survey responses will not be identified by name. The link to SSA electronic folder numbers and study identification number will be destroyed after the study completion and after the acceptance for publication, if appropriate. Any information respondents provide will be available only to research staff. All information respondents provide in this study will be only for research purposes and their name will not be used in any publication that may be written from this research. This research project will be conducted in compliance with all applicable state and federal laws.

The web survey system that Westat has developed for this project will not contain or be linked to claimant identifying information. Only Westat's unique identification number for each claimant along with his or her survey responses will be stored on Westat's system and the BU-CAT-SMS.

A Privacy Impact Assessment (PIA) has been completed for work related to this project **(Attachment 12)**.

This project has been approved by the Boston University Institutional Review Board **(Attachment 13)**. The NIH Institutional Review Board (NIH-IRB) granted a reliance agreement with Boston University covering projects relating to this IAA **(Attachment 14)**.

The NIH collaboration with SSA is conducted under the *NIH SORN 09-25-0200; system name: Clinical, Basic and Population-based Research Studies of the National Institutes of Health (NIH), HHS/NIH/OD,* published in the Federal Register on September 26, 2002.

## A.11 Justification for Sensitive Questions

Questions included in the CAT instrument short form for *Behavioral Health* are designed to assess a respondents functioning in interpersonal domains. This includes sensitive questions that can be regarded as "psychological problems" including questions about feelings towards others, and mood swings. These questions are essential to this study activity as they are included in the instruments, including the legacy instruments, which are widely used and will be compared to the BU-HDR CAT instruments (FABs). It

would be impossible to complete development of the FABs without including these questions.

While Polimetrix retains personally identifiable information (PII) for individuals who choose to participate in their surveys (name, address), the NIH along with its contractor, Boston University, will not be provided with that information. Boston University will be provided with de-identified data that only includes demographic information including: age, race, gender, marital status, education, and zip code.

Respondent consent will be obtained by Polimetrix. Polimetrix will retain responsibility and oversight of the consent process. Since all respondents will take the survey online they will review and click through a consent text form prior to being able to start the survey.

## A.12 ESTIMATES OF HOUR BURDEN INCLUDING ANNUALIZED HOURLY COSTS

Calibration

Of the 20,000 potential participants who receive a pre-notification package, it is estimated 7,800 claimants will complete initial screener interviews.

We will recruit 3,500 individuals to participate in the first survey at an estimated burden of 1 hour each (15 minutes for consent and 45 minutes to complete the survey). We estimate 3,000 individuals will complete the second screener, consent and survey. For both Survey 1 and 2, the burden required for the voluntary participants to listen to or read follow-up/reminder messages, if necessary, is factored into the burden calculations. In every instance, we assume the maximum possible burden per participant. For the predictive validity portion of this phase of study, participants will not be further contacted, and will incur no additional burden.

For the normative sample of 2,000 individuals, it is estimated that they will incur no more than 45 minutes of burden for each survey, including time to complete the consent documents.

The estimates of hour burden provided below (Table A.12-1) are based on the research experience during a similar item bank development project, which respondents completed similar types of items. These estimates are calculated for the calibration phase only. Once study design and materials are finalized for the remaining phases, necessary change requests will be submitted to account for any additional burden to conduct the remaining portions of this study.

**A.12-1**

| Type of Respondents | Number of Respondents | Frequency of Response | Average Time per Response (in hours) | Annual Hour Burden |
|---|---|---|---|---|
| Calibration Phase | | | | |
| Survey 1 -Screener Call (Not Interested) | 12,200 | 1 | 3/60 | 610 |
| Survey 1- Screener Call (Participate/Eligible) | 7,800 | 1 | 15/60 | 1,950 |
| Survey 1- Consent Form | 3,500 | 1 | 15/60 | 875 |
| SSA Claimant Survey 1 | 3,500 | 1 | 45/60 | 2,625 |
| Survey 2- Screener Call (Not Interested) | 500 | 1 | 3/60 | 25 |
| Survey 2- Screener Call (Participate Eligible) | 3,000 | 1 | 15/60 | 750 |
| Survey 2 – Consent Form | 3,000 | 1 | 15/60 | 750 |
| SSA Claimant Survey 2 | 3,000 | 1 | 45/60 | 2,250 |
| Normative Population Survey 1 | 2,000 | 1 | 45/60 | 1,500 |
| Normative Population Survey 2 | 2,000 | 1 | 45/60 | 1,500 |
| TOTAL | | | | 12,835 |

## A.12-2 ANNUALIZED COST TO RESPONDENTS

With respect to time costs, all SSA claimants completing the instruments will have indicated that they are "work disabled." We are therefore assuming, for the purposes of this validation study, that these individuals are currently not employed; however, for all samples in each phase of study, we have calculated time costs in table A.12-2 below.

The estimated total time cost for all respondents, in all phases of this study, is $353,547. This is based on a mean hourly wage of $24.45 reported by the Bureau of Labor Statistics in June 2014 (http:// www.bls.gov/oes/current/oes_nat)

| Type of Respondents | Number of Respondents | Frequency of Response | Average Time per Respondents | Hourly Wage Rate | Respondent Cost |
|---|---|---|---|---|---|
| Calibration Phase | | | | | |
| Screener Call (Not Interested) | 12,200 | 1 | 3/60 | $24.45 | $14,914.50 |
| Screener Call (Interested) | 7,800 | 1 | 15/60 | $24.45 | $47,677.50 |
| Consent Survey 1 | 3,500 | 1 | 15/60 | $24.45 | $21,393.75 |
| SSA Claimant Sample Contact 1 | 3,500 | 1 | 1 | $24.45 | $85,575 |
| Screener Call 2 (Not interested) | 500 | 1 | 3/60 | $24.45 | $611.25 |
| Screener Call 2 | 3,000 | 1 | 15/60 | $24.45 | $18,337.50 |
| Consent Survey 2 | 3,000 | 1 | 15/60 | $24.45 | $18,337.50 |
| SSA Claimant Sample Contact 1 | 3,000 | 1 | 1 | $24.45 | $73,350 |
| Normative Population Contact 1 | 2,000 | 1 | 45/60 | $24.45 | $36,675 |
| Normative Population Contact 2 | 2,000 | 1 | 45/60 | $24.45 | $36,675 |
| **Totals** | | | | | **$353,547** |

## A.13 ESTIMATE OF OTHER TOTAL ANNUAL COST BURDEN TO RESPONDENTS OR RECORD KEEPERS

There are no costs to respondents beyond time.

## A.14 ANNUALIZED COST TO THE FEDERAL GOVERNMENT

This study is being supported through a NIH contract with BU, Contract No. HHSN269201200005C: Computer Adaptive Tools (CAT) Development. It is estimated that the cost of subcontracting to Westat and YouGovPolimetrix for the data collection

portion of the validation, as well as the cost of Boston University and NIH/CC/RMD research staff will cost $2,179,419.

| **A.14 - 1 Estimate of Annualized Cost to the Federal Government** | |
|---|---|
| Total costs for Westat through subcontract with BU, including labor costs and all related data-collection activities | **$1,740,104** |
| Total cost for YouGov Polimetrix through subcontract with BU, including labor cost and all related data-collection activities | **$90,500** |
| Portion of Contract Costs for Boston University Personnel to support information collection and analytic work | **$312,704** |
| Staff time costs at NIH/CC/RMD | **$36,111** |
| *Principal Investigator:*<br>*Title 42/ $250,000 per year/ 5% effort* | $12,500 |
| *Protocol Manager (project oversight):*<br>*GS 12 Step 5/ $84,860/ 10% effort* | $ 8,486 |
| *Management Analyst (regulatory requirements):*<br>*Contractor/ $75,625/ 20% effort* | $15,125 |
| **Estimate of Annualized Costs to the Federal Government:** | **$2,179,419** |

Estimates for Polimetrix costs are based on market research and the terms of the awarded contract. Contract costs and FTE costs are estimated based on hours of personnel support required to complete data analysis and project management.

## A.15 EXPLANATION FOR PROGRAM CHANGES OR ADJUSTMENTS

Not applicable, this is a new collection of information.

## A.16 PLANS FOR TABULATION AND PUBLICATION AND PROJECT TIME SCHEDULE

This study is part of a larger multi- year scientific project focused on improving SSA's disability determination process. Data collection is projected to begin in August of 2014 and completed by October of 2014. Subsequent data analysis should be completed by March of 2015. This work is part of an existing contract with BU. These data will be analyzed and outcomes published as part of the larger project work.

The data analysis will address the following parameters:

- Response burden
- Score precision
- Internal consistency & reliability
- Score range (ie.,floor or ceiling effects)
- Predictive validity

To monitor the BU-HDR FABs in real time, we will calculate the standardized log-likelihood statistic ($l_z$) for polytomous items to test the person fit. The empirical distribution of the log-likelihood statistic is reasonably close to a standardized normal distribution, so we will calculate the percentage of respondents in which $l_z$ exceeded an alpha level of .05.

Response burden will be measured as the average amount of time it takes to complete instrument. A t-test will be used to assess whether the average amount of administration time between the BU-HDR FABs and other measurements is significantly different.

To illustrate the difference in precision in score range across instruments, we will calculate the average Standard Error (SE) along the entire scale continuum across different instruments. We will use the t-test to assess whether the average SE is significantly different between BU-HDR FABs and other measurements at different score ranges.

To examine internal consistency, we will use marginal reliability calculations that are specific to item response theory (IRT) which allow us to compare BU-HDR FABs with other instruments. Marginal reliabilities are similar to Cronbach's alpha coefficient used in classical measurement theory in that it is a measure of how well items within a domain relate to each other.

The percentage of ceiling and flooring will be calculated in each instrument. A chi-square test will be used to test whether the percentages of ceiling or flooring are significant different between BU-HDR FABs and other instruments.

**A.16 – 1: Project Time Schedule**

| Activity | Time Schedule |
|---|---|
| Invitation (pre-notification packages) to Claimants from Westat | 1 - 2 weeks after OMB approval |
| Online data collection | 0.5 - 1 month after OMB approval |
| Analyses | 4-6 months after OMB approval |
| Predictive Validity Determined | 25 months following OMB approval |

| | |
|---|---|
| Study Activity Complete | 36 months after OMB approval |

## A.17 Reason(s) Display of OMB Expiration Date is Inappropriate

Not applicable, the expiration date will be displayed on all data collection instruments including web-administered surveys.

## A.18 Exceptions to Certification for Paperwork Reduction Act

No exceptions are requested.