Supporting Statement B for

The SSA-NIH Collaboration to Improve the Disability Determination Process: Validation of IRT-CAT tools
NIH/CC/RMD

Name: Meghan Gleason
Address: 10 Center Drive, Rm 1-2420
Telephone: 301-443-9085
Fax:
Email: Meghan.gleason@nih.gov

Table of contents

B.1 Respondent Universe and Sampling Methods

The information from the proposed data collection will be used by the NIH Clinical Center (through a contract with Boston University and sub-contract with YouGovPolimetrix (YGP), a survey research firm based in Palo Alto, CA) to validate the CAT instruments. The proposed information collection will support psychometric testing. Specifically, the validation will seek to address three aims or research questions:

Aim #1: What are the psychometric properties of the SSA-CATS compared to legacy instruments?

The data analysis will address the following psychometric parameters:

- Score precision
- Internal consistency reliability
- Score range (ie., floor or ceiling effects)

To monitor the SSA-CATs 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 subjects in which l_z exceeded an alpha level of .05. We will then test the following psychometric parameters:

Precision: 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 SSA-CATs and other measurements at different score ranges.

Reliability: To examine internal consistency, we will use marginal reliability calculations that are specific to item response theory (IRT) which allow us to compare SSA-CATs 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.

Score range: 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 SSA-CATs and other instruments.

Aim #2: What is the response burden of the SSA-CATs compared with legacy instruments?

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 SSA-CATs and other measurements is significantly different.

Aim #3: Do the SSA CATs measure the underlying concept(s) that we purport they are measuring?

To assess the SSA CATs validity, we will analyze the concurrent validity of the SSA-CATs and selected legacy measures using Pearson correlation coefficients. Specifically, Pearson correlations coefficients will be calculated between scores from the SSA Physical Capabilities CAT and the SF-36 scale scores, and the PROMIS Physical Function CAT scores; between the SSA Interpersonal Interactions CAT and the SF-36 scale scores, and the BASIS-24 scale scores.

Respondents will be recruited through Polimetrix, which recruits for studies using an opt-in panel of 1.5 million U.S residents who have agreed to participate in Polimetrix's Web surveys. Panel members are recruited by a number of methods to help ensure diversity in the panel population. Recruiting methods include Web advertising campaigns (both text and banners), permission-based email campaigns, partner sponsored solicitations (*e.g.*, Rock the Vote and Cox Communications), telephone-to-Web recruitment, and mail-to-Web recruitment. By utilizing different modes of recruitment continuously over time, this ensures that hard-to-reach populations will be adequately represented in survey samples. Participants are not paid to join the PollingPoint panel, but do receive modest incentives through a loyalty program to take individual surveys.

Polimetrix tracks employment status within their active participant pools. They currently have about 5,600 "permanently disabled" participants which will serve as the sampling frame for this pilot study. YouGov will exclude any permanently disabled respondent who participated in the normative calibration study from participating in the validation study. From this population, Polimetrix will recruit 1,000 participants to answer 70-86 items if they claim a primary physical impairment and 88-96 items if they claim primary mental health impairment.

Sampling Methods:
Conventionally, one would then attempt to contact the respondents in a 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 listed phone numbers – and those that do may not agree to participate. The permanently disabled sample in the validation study is not anticipated to be matched to the respondents in the normative study. However, if the sample is matched, it would not cause any problems of analysis or interpretation.

Instead, for each member of the target sample, Polimetrix will select one or more matching members from their pool of opt-in respondents. This pool has been recruited by a variety of means (banner ads, email lists, promotions and offers). 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. A “usable samples” is defined as individuals with demographic characteristics representative of a target population or group. Polimetrix uses a model that applies a matching strategy to their vast opt-in pool of potential respondents. These respondents are drawn from a group of people who volunteered to do surveys and thus are not necessarily representative in their attitudes (responses) in the same way that a random sample of respondents from the target population would be. It is important to remember that the appropriate comparison is not with the results obtained from a hypothetical random sample from a complete sampling frame with a 100% response rate. Rather, any other approach by phone or mail or e-mail is going to yield a sample with some self-selection bias as well at much greater cost, assuming that one could even get a complete sampling frame of the target population, involving no errors of omission or commission in the listing of elements to be sampled, in the first place. In addition, the Polimetrix model has the advantage that the people being contacted are not citizens who are being imposed upon unexpectedly or unwillingly by the survey process.

Matching is done on a large set of variables available in both the population enumeration database and the op-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. Various types of matching are possible (e.g., exact matching, propensity score matching); Polimetrix employs a proximity matching method whereby a distance function is computed for each attribute (e.g., age, years of schooling, latitude and longitude of residence) 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.

The active participant pool of "permanently disabled" participants is about 5,600 individuals and their demographic information is provided in Table 1. The total number of participants differs somewhat across demographic characteristics (from 5656 to 5677) due to small amounts of missing data (.4% maximum missing).

Table 1. Demographics of “Permanently Disabled” Participants

Gender	#	%
Men	2679	47.2%
Women	2998	52.8%
Total	5677	100%
Race		
White	4660	82.1%
Black	341	6%
Latino	123	2.2%
Other*	553	9.7%
Total	5677	100%

Education		
High school or less	1613	29%
Some post-HS**	2813	49.7%
College	839	14.8%
Post Graduate	400	7.1%
Total	5665	100%
Region		
Northeast	930	16.4%
Midwest	1323	23.4%
South	2137	37.8%
West	1266	22.4%
Total	5656	100%

*Includes Asian, Native American, Middle Eastern, Mixed and Other.
** Includes "some college" and 2-year degree graduates.

The validation study will involve individuals in the sub-group of the overall YouGov pool of voluntary opt-in survey respondents who have self-reported their employment status as permanently disabled. For purposes of the validation study, we anticipate sufficient variation on level of disability within this group because (1) it is relatively large (the YouGov pool of "permanently disabled" participants has increased by more than 50% since submission of this application; Table 1 reflects updated figures) and (2) diverse with regard to gender (47% male, 53% female), racial/ethnic background (18% minority), education (29% high school or less), and geographic distribution (Northeast 16%; Midwest 23%; South 38%; West 22%). Given this and the substantial size of the target sample, we expect that their scores on the both the legacy and new CAT-based scores will exhibit sufficient variation to allow for a reasonable assessment of the strength of their relationship.

B.2 Procedures for the Collection of Information

Polimetrix recruits for studies using an opt-in panel of 1.5 million U.S residents who have agreed to participate in Polimetrix's Web surveys. Panel members are recruited by a number of methods to help ensure diversity in the panel population. Recruiting methods include Web advertising campaigns (both text and banners), permission-based email campaigns, partner sponsored solicitations (*e.g.*, Rock the Vote and Cox Communications), telephone-to-Web recruitment, and mail-to-Web recruitment. By utilizing different modes of recruitment continuously over time, this ensures that hard-to-reach populations will be adequately represented in survey samples. Participants are not paid to join the PollingPoint panel, but do receive modest incentives through a loyalty program to take individual surveys.

Polimetrix tracks employment status within their active participant pools. They currently have about 5,600 "permanently disabled" participants which will serve as the sampling frame for this pilot study. From this population, Polimetrix will recruit 1,000 participants to answer 70-86 items if they claim a primary physical impairment and 88-96 items if they claim primary mental health impairment. The participants will be matched with our calibration sample on age, gender, race, and education. Study participants will be asked a screener question if the reason for their "permanently disabled" employment status is the result of a primary physical or mental health impairment. This information will be used to match each potential subject to the appropriate CAT content domain (ie., Physical Demands or Interpersonal Interactions). If the study participant claims dual impairments, they will be placed into the group requiring more study completes.

All of the Polimetrix panelists have provided their e-mail so that they may receive survey invitations to participate in surveys. As a policy, no one panelist is invited to take more than 12 surveys in a year. Additionally, with each survey invitation they are reminded of the Polimetrix policy on privacy, the opportunity to immediately opt-out, and of the voluntary nature of each request regardless of the survey sponsor.

Measures for Assessing Workplace Physical Function Demands

Study participants who indicate that their primary reason for not being able to work is the result of a physical impairment be asked to complete the SSA Physical Demands CAT tool, the SF-36, and the PROMIS Physical Functioning CAT. The SF-36 is a widely used multi-purpose, short-form survey with 36 questions that measure physical and mental health. It has been broadly tested in general and disease specific populations. The SF-36 consists of eight scaled scores: vitality, physical functioning, bodily pain, general health, physical role functioning, emotional role functioning, social role functioning, and mental health. We selected the SF-36 as a legacy instrument because of its extensive history and use in research and the content coverage it provides. The PROMIS Physical Function item pool consists of items covering activities of daily living, lower extremity, and central body functions. PROMIS utilizes rigorous methodology for developing its measures and testing their validity. This work integrates qualitative and quantitative research and psychometrics. Content and disease experts as well as thousands of patients provided input into the development process. The PROMIS item pools have been tested and validated in clinical and generic populations. The PROMIS physical function CAT was selected because of the content coverage it provides as well the extensive and rigorous testing process the PROMIS initiative utilized.

Measures for Assessing Workplace Interpersonal Interaction Demands

Study participants who indicate that their primary reason for not being able to work is the result of a mental health impairment will be asked to complete the SSA Interpersonal Interaction CAT tool, the SF-36 and the Behavior And Symptom Identification Scale (BASIS-24.) The BASIS-24 is a leading behavioral health assessment tool. The BASIS-24 underwent extensive field testing as part of a multiyear research and development process. The survey was tested on more than 6,000 participants from racially and ethnically diverse backgrounds receiving inpatient or outpatient treatment for mental

health or substance abuse at one of 28 facilities across the U.S. The development of the survey was grounded in Item Response Theory (IRT) methods. In order to comply with SSA requests we will remove 4 items from the BASIS-24 that ask about alcohol and drug use.

To address order effects during test administration, we are planning to "counter-balance" the mode of administration by randomly assigning half the sample take the SSA-CAT first and half the sample will take the legacy items first. For both study groups, it should take approximately 30 minutes to complete the assessments and the assessments will only be collected once.

Table 1. Summary of Survey Content in Two Domains

Domain	BU-HDR CAT tool	SF-36	PROMIS PF CAT	BASIS-24	Total Per Domain
Physical Function	24-30 items	36 items	10-20	0	70-86 items
Interpersonal Interactions	32-40 items	36 items	0	20 (removed 4 alcohol/drug items)	88-96 items
Total	56-70 items	72 items	10-20 items	20 items	

B.3 Methods to Maximize Response Rates and Deal with Nonresponse

Polimetrix 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. Polimetrix's response rate is estimated at about 70%. This response rate is estimated by Polimetrix through tracking respondent rates for other surveys they conduct. It is important to note that all individuals who will be contacted with respect to this data collection will have already volunteered to participate in Polimetrix's opt-in survey panel, and therefore represents a low burden on respondents and high likelihood of achieving statistically required completed surveys. Polimetrix will not re-contact any individual who has refused to participate in this survey.

B.4 Test of Procedures or Methods to be Undertaken

The data analysis will address the following parameters:

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

- Concurrent validity

To monitor the BU-HDR CAT 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 subjects 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 CAT 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 CAT 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 CAT 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 CAT and other instruments.

We will analyze the concurrent validity of the BU-HDR CATs and other measures using Pearson correlation coefficients. Specifically, Pearson correlations coefficients will be calculated between scores from the BU-HDR physical function CAT and the SF-36 scale scores, and the PROMIS Physical Function CAT scores; between the BU-HDR interpersonal interactions CAT and the SF-36 scale scores, and the BASIS-24 scale scores.

While correction for attenuation of correlation is possible, this has not been a focus in the literature with respect to development of health-related IRT/CAT instruments. Attenuation of correlation refers to the understanding that any observed relationship between two scales is less than it "really" is because measurement error obscures the 'true'relationship; without error, it would be stronger. This makes sense if one thinks of any given score as decomposable into true scores plus error. The error components of two scores being correlated should be unrelated, and so the larger those error components are, the more the observed relationship will be attenuated/masked. From the theory of measurement error, it is possible to actually estimate how much effect measurement error has on a particular correlation between two measures – i.e., how much the correlation between "true scores" would be higher than those between the observed fallible scores. Nunnally observes that the formulate for making this estimate: 'corrected' correlation ("..really an estimate of how much the correlation would be if two variables were made perfectly reliable") =

(1) Numerator: observed correlation between A and B
(2) Denominator: [Sq root of reliability of A] x [Sq root of reliability of B]

Journal articles in our field do not typically make reference to the correction for attenuation or its implications, so we are not planning to do so for this validation study.

Ref: J.C. Nunnally. Psychometric Theory (2nd edition). McGraw Hill: 1978.

We are not proposing to pilot test this study prior to the start of data collection. BU-HDR has demonstrated the ability to design and administer large-scale web-based calibration studies. For more than a decade, Drs. Haley, Jette, and the BU-HDR team have developed numerous items banks and successfully conducted numerous calibration studies leading to the development and dissemination of health related CATs. The BU-HDR team has developed both on-site and off-site recruiting procedures for calibration studies. We have been very successful in recruiting the projected numbers for all of the calibration projects. Polimetrix has an extensive track record of collaboration with academic and research institutions in the area of health and functional status assessment in general, and BU-HDR has previous experience working with Polimetrix on several studies.

B.5 Individuals Consulted on Statistical Aspects and Individuals Collecting and/or Analyzing Data

Individuals Consulted		
Pengsheng Ni	Boston University Boston, MA	617-638-1989
Alan Jette	Boston University Boston, MA	617-638-1985
Mark Meterko	Boston University Boston, MA	(857) 364-4433

Individuals Collecting and/or Analyzing Data		
YouGovPolimetrix Inc. (YGP)	Palo Alto, CA	Collect
Mark Meterko	Boston University Boston, MA	Analyze
Pengsheng Ni	Boston University Boston, MA	Analyze

Updated: 12/04/2007 by: MPC