TeleSage Outcome Measurement System (TOMS):
An Overview of Item Bank Development and Psychometric Properties

About TeleSage. Inc.

TeleSage, Inc. is a privately owned, for-profit company located in Chapel Hill, NC. Founded by Benjamin Brodey, MD, MPH, a graduate of MIT and Harvard Medical School, TeleSage has been the recipient of 7 NIH Phase II SBIR grant awards, totaling over six-million dollars, and has pioneered the development and application of automated survey administration and clinical reporting technologies for mental health and substance abuse treatment and clinical research.

Dr. Brodey and his team have many years of experience tracking the patient-centered behavioral health outcomes of patients receiving both publicly and privately funded behavioral health services. TeleSage has administered over 350,000 longitudinal behavioral health outcomes tracking surveys. TeleSage has participated in statewide outcomes tracking initiatives in California, Ohio, Idaho, Iowa, Tennessee, and Washington as well as consulted with a number of other states developing outcomes tracking systems. In addition, TeleSage has sold its software to hundreds of research organizations, including over 30 research universities.

Dr. Brodey has substantial experience with outcomes tracking initiatives. He was a consultant to SAMHSA on the development of the outcomes tracking system used in the Decision Support DS2000+) project. Dr. Brodey has also developed an ongoing relationship with the NIH Patient-reported Outcomes Measurement Information System (PROMIS) project. He was an invited presenter at the PROMIS steering committee in April 2006 and served as a presenter and moderator at additional PROMIS conferences.

Item Development

TeleSage engages in a multi-step item development process for all behavioral health symptom and functioning domains. This process includes rigorous procedures for item authoring, cognitive interviewing with intended populations, Item Response Theory (IRT) and Differential Item Functioning (DIF) statistical analyses, comparison with gold-standard behavioral health assessments, and assessment of consumer satisfaction, where applicable.

Item authoring.

The first step in the TeleSage item development process is the thorough review of existing instruments for core concepts related to symptoms and functioning. Where applicable, core concepts are also obtained from the Diagnostic and Statistical Manual for Metal Disorders (DSM). For example, when developing the depression item bank, the BASIS-32, BDI-II, CES-D, Geriatric Depression Scale – 10, HAM-D, Prime-MD and PHQ-9, SF-36, the Zung Depression Scale, among many other instruments, as well as criteria in the DSM-IV-TR for Major Depressive Episode were reviewed for core concepts.

The second step in the item development process is the creation of survey items. New, concise items are written that encompass key concepts and sub-domains. Specifically, items are written that are free from lead phrases (e.g., In the morning,…), contingent questions (e.g., I felt upset for no reason), multiple concepts (e.g., I feel more restless or wound up than usual), and idiomatic/culture-specific language (e.g., Things have been getting on top of me). In several instances, more than one version of a question is possible and therefore created. All items use a 5-point Likert response set (*never, rarely, sometimes, often, always*[1]) and “during the past 7 days” or “during the past 30 days” timeframes, as appropriate.

The third step in the item development process is the review of the item pool by a panel of experts in eh appropriate content area. Each member of the expert panel independently reviews the item pool and evaluates the items for clarity, as well as how well each item represents the core concepts. Items evaluated as both clear and representative are retained, while items that are unclear or peripheral are omitted or rewritten, as appropriate. These procedures result in a final item pool of items to be tested during cognitive interviewing. The resulting items are all evaluated to have face-validity by our panel of experts, as a result of this rigorous set of item development procedures.

Cognitive interviewing.

Cognitive interviewing (CI) has been used most often to aid instrument development and has been successfully employed in the development of health-related quality of life (HRQOL) and patient-reported outcomes (PRO) assessments (Willis, Reeve, & Barofsky, 2004). Cognitive interviewing refers to a combination of techniques that fall into two general categories: thinkalouds and verbal probes (DeMaio & Rothgeb, 1996; Willis et al., 1999). In a thinkaloud interview, the interviewee verbalizes his or her thoughts while engaged in a cognitive activity, such as responding to an item on a questionnaire, with little interjection by the interviewer. Thinkaloud interviews can be conducted while the cognitive process is occurring (concurrent thinkaloud) or after the cognitive process has been completed (retrospective think aloud) (Sudman et al., 1996). Verbal probes are questions that ask participants about their cognitive processes. Due to the qualitative nature of cognitive interviews, a sample size of 5 to 15 participants is typical per interviewing round (Willis, 2005).

As part of the item development process, cognitive interviews are conducted with a diverse sample of behavioral health consumers who are recruited from both public and private behavioral health clinics. Consumers represent a range of diagnoses, race, ethnicity, and gender. Each participant is presented with ten questions at a time. The participant is asked to engage in a thinkaloud while reading through the items and responding to them so that the interviewer can get a sense of the participant’s thought process. Once the participant completes the thinkaloud for the 10-items, the cognitive interviewer goes back through each question and asks specific questions about each question to better understand any difficulty the participant had understanding or answering the question. The cognitive interviewer then proceeds with the next set of ten questions using the same procedures until all items are administered or the session time runs out. The cognitive interviewer takes notes during the session indicating whether the participant understood the item, understood the item differently than it was intended, and/or asked for clarification about an item. Questions are administered in a partially randomized order in order to ensure a large enough sample size per question. As a result, all items are typically responded to by at least half of the participants within each population type, depending on the number of questions. Cognitive interviewing is usually completed in three rounds, with approximately 15-20 participants in the first round. Questions that pose difficulties for more than 20 percent of participants are deleted from the item pool or revised and retested. The second and third rounds are required when questions have been revised and need to be tested. Subsequent rounds of interviewing usually involve 5-15 participants per round, depending on the number of questions.

Patient Sample

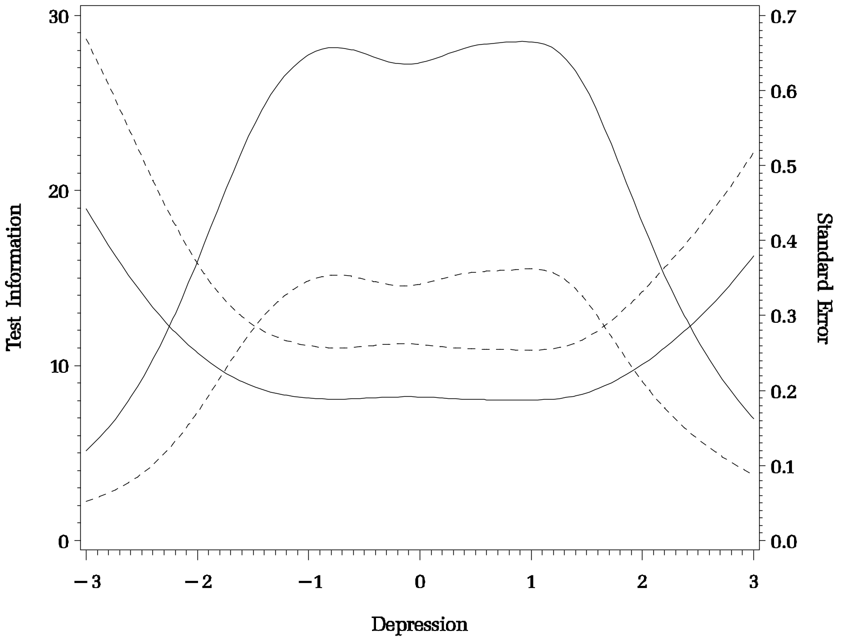
As part of their NIH-funded research, TeleSage recruits large samples from both public and private behavioral health clinics (typically 300-600 from each clinic type), and a community sample, as needed, in order to establish the validity and reliability of its item pools. Each sample represents a diverse population in terms of race, ethnicity, gender, and age.

Reliability and Validity

Item Response Theory (IRT) Analyses.

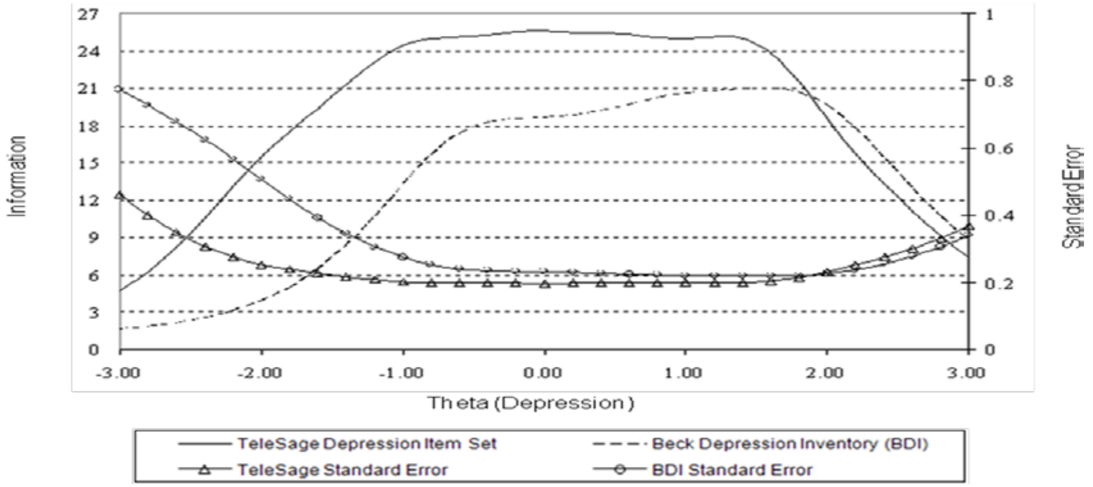
TeleSage conducts both Classical Test Theory (CTT) and Item Response Theory (IRT) analyses to establish reliability. A clear advantage of IRT compared to CTT is the ability to estimate the Test Information Function (TIF) over the range of a construct. The TIF, in essence, provides a graphical representation of how well the test (or scale) is measuring the construct at any given level of the construct. Higher information values correspond to smaller standard errors and greater confidence in individual scores. As illustrated in Figure 1, the TIF for the Brodey Depression Scale, an assessment utilizing a subset of the TeleSage depression items, is high across a wide range of depression. More specifically, the TIF suggests that we can have confidence in the measurement of individuals with depression levels ranging from approximately -2 S.D. below the mean to 2.5 S.D. above the mean level of depression in the study population. The measure performs best from approximately -1 S.D. below the mean to 1.5 S.D. above the mean level of depression. TIFs for all other domains are similar in that the TeleSage items for other domains perform well across a range of severity within each domain.

Figure 1. Test Information Function for the Brodey Depression Scale, long (solid line) and short (dashed line) forms



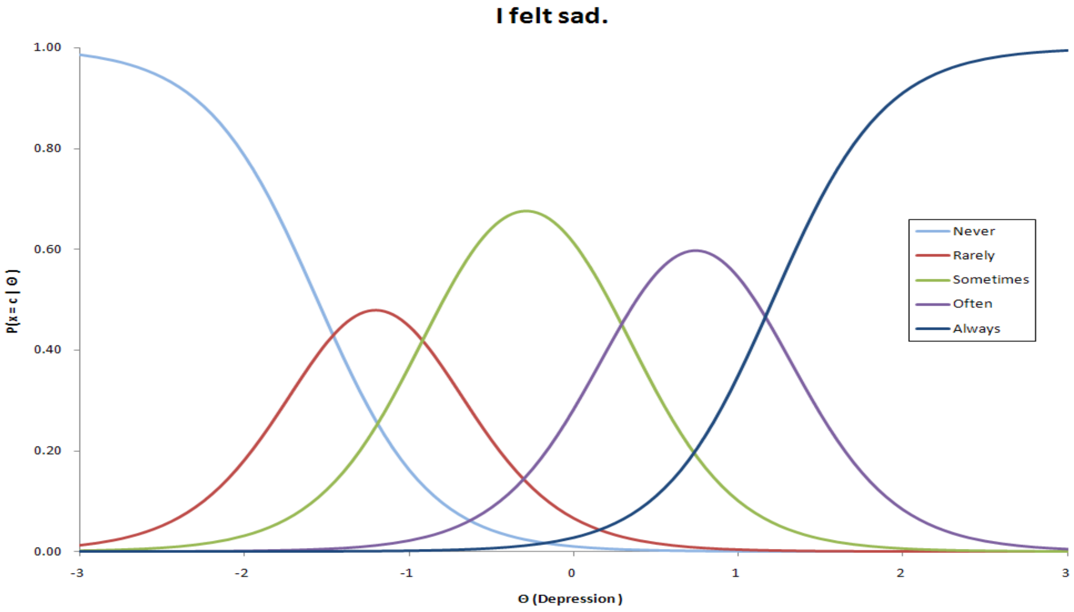
In addition, a comparison of the TIF for the TeleSage depression domain with the TIF for the Beck Depression Inventory (BDI-II), illustrated in Figure 2, demonstrates the strengths of the newly developed item bank. Our items differentiate people better than the BDI-II at low and moderate levels of depression. However, the BDI-II has a very slight advantage over TeleSage depression items at the high extreme of depression.

Figure 2. Test Information Function for the Brodey Depression Scale and the Beck Depression Inventory



Item Information Curves (IIC) are also graphed for each item, illustrating the probability that a respondent will provide each of the five response options at a given level of theta (NOTE: theta is a precise measure of the domain of interest, e.g., depression).

Figure 3. Item Information Curve for a Sample Item from the TeleSage Depression Item Bank



Preliminary polytomous IRT analyses for all domains found that the items performed well (i.e., fit the data well and demonstrated clear discrimination).

Differential Item Functioning.

TeleSage items are also examined for Differential Item Functioning (DIF) for gender, SES, and race. For example, of 25 core depression items, 14 items were found to exhibit no DIF. Few meaningful differences were detected among the additional 11 items, but when meaningful DIF occurred, individual item parameters were estimated for each sample of interest. Results generally indicate that the TeleSage item banks are sensitive enough to detect differences in symptom severity and functionality, while not exhibiting race, gender, or clinic site (public vs. private) differences. Therefore, the TeleSage item banks can be successfully used in both public and private clinic sites with males and females, both Caucasians and non-Caucasians.

Test-Retest Reliability.

Test-retest reliability of the adult TeleSage item banks, including two additional domains: Psychoticism and Recovery, was established using a sample of public and private behavioral health clinic outpatients. Participants completed the items twice, approximately 4-7 days apart. There were no significant differences in the test-retest reliability of the items between the private and public behavioral health clinics. A high reliability coefficient (*r*=.76) was found for Symptom domains and a high reliability coefficient (*r*=.84) was found for Functioning domains. These results provide evidence that the TeleSage adult item banks are reliable for use in both public and private behavioral health clinics.

Convergent & discriminant validity.

To establish validity, 'gold-standard' instruments corresponding to key domains are administered alongside the TeleSage item pools. For example, during validation of the adult Depression domain, the Beck Depression Inventory-II (BDI-II), the nine-item Patient Health Questionnaire (PHQ-9), and the PROMIS depression long scale, were administered to establish convergent validity with the depression domain, but discriminant validity for the other domains. Correlations between the final domains, selected after IRT analyses, were computed and reported.

Table 1. Sample Correlations Between the BDI-II and Sample TeleSage Item Bank Domains

	BDI-II
Depression	**.83**
Anxiety	.78
Anger	.49
Psychoticism	.42
Recovery	-.62
Work Functioning	-.54
Social Functioning	-.70

*All results significant at *p*<.0001.

Consumer Satisfaction with Administration Modality

When test-retest reliability was assessed with the TeleSage item banks, consumer satisfaction with administration modality (computer touchscreen, Interactive Voice Response (IVR), or Personal Digital Assistant (PDA)) was also assessed. Overall, no significant difference was found for preference of taking the survey on a computer touchscreen compared to a PDA, whereas both modalities were preferred over IVR (phone) surveys. Further, consumers rated the touchscreen as the easiest modality to view the questions on, but no significant difference was detected among modalities for ease of understanding the survey questions.

Conclusions

Overall, TeleSage item banks:

1) Have clinical relevance, as determined by expert review and cognitive interviewing with behavioral health consumers.
2) Correlate in expected ways with 'gold-standard' instruments, and therefore have established convergent and discriminant validity.
3) Contain items with excellent psychometric properties according to IRT analyses, and are therefore reliable.
4) Are largely free from meaningful Differential Item Functioning (DIF) for clinic type, race, ethnicity, and gender.
5) Contain items with separate established item parameters per sample when meaningful DIF exists.
6) Are reliable according to results from IRT analyses and test-retest administration.
7) Can be successfully administered via paper, computer touchscreen, phone (IVR), or PDA.

Related Publications

(please email bb@telesage.com to request copies)

Azocar, F., Cuffel, B., McCulloch, J., McCabe, J., Tani, S., & Brodey, B.B. (2007). Monitoring patient improvement and its relation to treatment outcomes in a managed behavioral health organization. *Journal for Healthcare Quality, 29*, 4-12.

Brodey, B.B., Cuffel, B., McCulloch, J., Tani, S., Maruish, M., Brodey, I., & Unützer, J. (2005). The acceptability and effectiveness of patient-reported assessments and feedback in a managed behavioral health care setting. *The American Journal of Managed Care, 11*, 774-780.

Brodey, B. B., McMullin, D., Kaminer, Y., Winters, K., Mosshart, E., Rosen, C., & Brodey, I. (2008). Psychometric characteristics of the revised T-ASI self-report instrument. *Substance Abuse: Journal of the American Medical Education and Research in Substance Abuse*, *29,* 19-32.

Brodey, B.B., Rosen, C.S., Brodey, I.S., Sheetz, B.M., & Unützer, J. (2005) b. Reliability and acceptability of automated telephone surveys among Spanish- and English-speaking mental health services recipients. *Mental Health Services Research*, *7*, 181-184.

Brodey, B.B., Rosen, C.S., Winters, K.C., Brodey, I.S., Sheetz, B.M., & Steinfeld, R.R., et al. (2005). Conversion and validation of the Teen-Addiction Severity Index (T-ASI) for Internet and automated-telephone self-report administration. *Psychology of Addictive Behaviors*, *19*(1), 54-61.

Brodey, B.B., Rosen, C.S., Brodey, I.S. (2004). Validation of the Addiction Severity Index (ASI) for internet and automated telephone self-report administration. *Journal of Substance Abuse Treatment, 26*(4), 253-259.

Brodey, B. B., Wirth, R. J., Elliott Wilson, M., Ayer, D., Brooks-DeWeese, A., Stonerock, G. L., Koble, J., & Brodey, I. (2010). Development of the Brodey Depression Scale Using Item Response Theory. *Manuscript under development.*