Reliability Test of a Health Risk Assessment Designed for People with Developmental Disabilities

James W. Conroy, Ph.D.
Center for Outcome Analysis
426-B Darby Road
Havertown, PA 19083
jconroycoa@gmail.org
www.eoutcome.org

Abstract

This article reports a reliability study of a health risk screening tool for people with developmental and intellectual disabilities. It examines two of the three classical forms of reliability: internal consistency and test-retest reliability. The internal consistency findings were good, in the range of .828 to .836. Though the design for the test-retest part of this study did not fit the traditional small sample "artificial" experiment, the "real world" data indicated strong reliability, in the range of .921 to .958. Limitations on the design and suggestions for continued examination of the reliability and validity of this potentially important tool were discussed.

Monitoring and maintaining optimal health among vulnerable populations, particularly people with developmental and intellectual disabilities, is an increasingly important issue (Lewis et al., 2002). America has shifted sharply away from congregate, segregated, large scale institutions for people with developmental disabilities (Lakin, Larson, Salmi, & Webster, 2010). The new decentralized community based networks of small residential settings introduces a new challenge for health care providers: because the people are no longer "all in one place," it is necessary to create new ways to manage the need for examinations, information, and attention to risk factors (Sullivan et al., 2006).

A concrete example of this phenomenon of decentralization in the developmental disabilities field is provided by the deinstitutionalization of the Pennhurst Center in Pennsylvania. In 1978, a federal District court determined that conditions at the Center were so inhumane that they constituted violations of the civil rights of the people who lived there – including the Constitutional provision of the eighth amendment prohibiting cruel and unusual punishment (Halderman v. Pennhurst, 1977). This case was then heard by the U.S. Supreme Court three times.[1] Though the Supreme Court never fully affirmed the decisions of the lower courts, the case was a milestone in the disability civil rights movement, largely because it resulted in the movement of more than 1,000 people from one large institution with centralized health care to about 300 separate care and support homes and apartments in regular neighborhoods all across the region. The people then relied for their health care on generic community medical capacity plus Medicaid, Medicare, and private insurance plans.

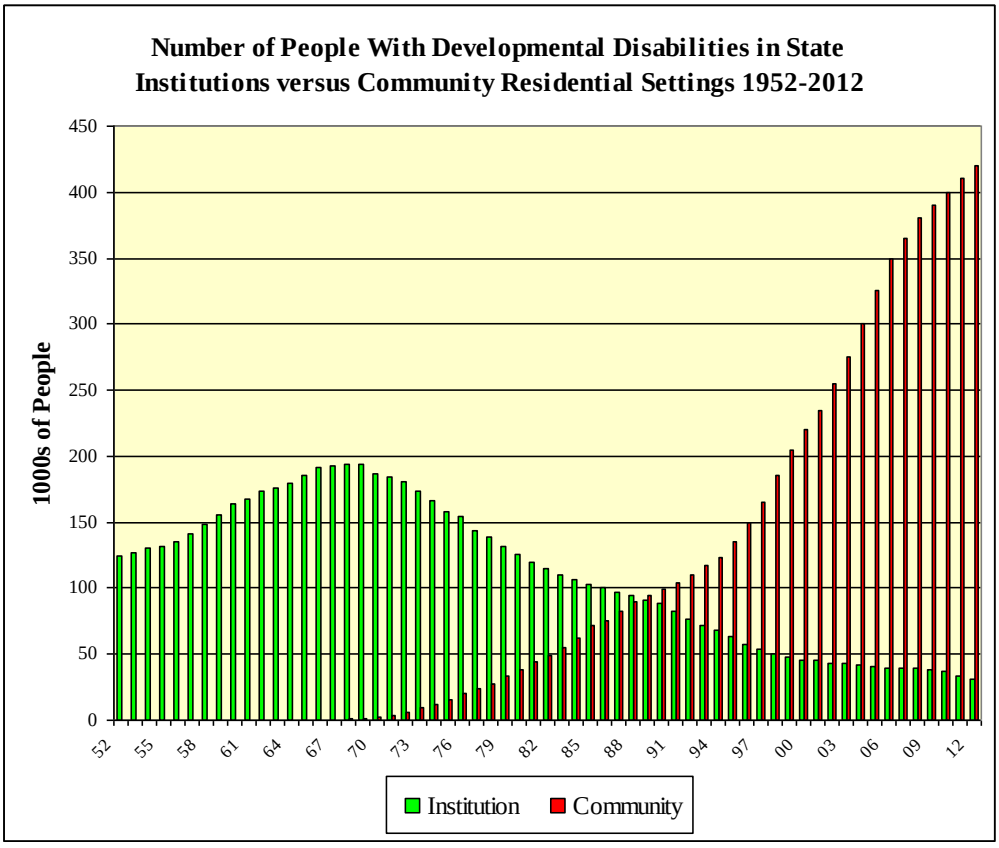
There was significant fear that the people so dispersed would be lost, neglected, homeless, or worse. This fear was largely based on the scandalous results of deinstitutionalization of America's state psychiatric hospitals for people with mental illness (Bassuk & Gerson, 1978). The Pennhurst Longitudinal Study (Conroy & Bradley, 1985) dispelled these fears with compelling findings that the people with developmental disabilities, in contrast to mental illness, who moved to communities were "better off" in nearly every way amenable to measurement. This included longevity (Conroy & Adler, 1998). This original study of deinstitutionalization of people with developmental disabilities was then replicated by dozens of studies in other states with extremely consistent findings (University of Minnesota, 2011).

This decentralization was then mirrored across all the states. In 1969, there were about 190,000 Americans with developmental disabilities in state institutions. Today,

[1] Halderman v. Pennhurst was recently cited by Justice Roberts no less than three times, in his decision for the majority finding the Affordable Care Act (Obamacare) to be constitutionally sound.

there are fewer than 35,000. This national change in systems and policy is represented in Figure 1.

Figure 1: Recent Patterns of Deinstitutionalization in America



The left side of Figure 1 shows the decline of the population state institutions for people with developmental disabilities, and the right side shows the massive increase in the population of small community homes. The graph shows visually how sharply American policy has changed since 1970.

The difficulty of providing health care to people so widely dispersed in this process was originally thought of as a major barrier. Fears that people would not receive the health services they needed were widespread. However, these fears were somewhat dispelled by an early study of the adequacy of health care in this decentralized community care system (Nowell, Baker, & Conroy, 1989). The study found that community health care was, indeed, initially inefficient and fragmented. Yet in most ways, even the fragmented community care was superior to the truly abysmal care provided previously at the institution. Community service providers had to work very hard to find health care for people the system had very rarely seen before, most of whom were quite healthy, but most of whom had no verbal communication. This made detection

and treatment of acute conditions unusually difficult. In later years, ongoing monitoring of the Pennhurst class members revealed that providers had evolved, learned, and made good health care routine (Conroy, 1993).

In spite of the clear fact that deinstitutionalization of Americans with developmental disabilities was a highly successful social policy, genuine concerns about monitoring health and health risk in a decentralized system do remain. Although most of the people who have moved have medical needs similar to those of the general population, some have biological insults and conditions that do pose higher than normal risks and high degrees of medical needs.

One approach to the new challenge of decentralization is to collect reliable data for each individual, and train all concerned workers – both health care and direct care – to recognize and report simple signs of potential import. That way, the clinical and non-clinical support workers who have contact with the residents in the thousands of widely distributed homes can be trained to complete simple, reliable screening tools that will enable accurate identification of high health risk.

This trend is already well under way in the generic health care field. The concept is now entrenched to the degree that it has a full heading in Wikipedia as "Health Risk Assessment" or HRA. (http://en.wikipedia.org/wiki/Health_risk_assessment). According to Alexander (2000),

> A health risk assessment (HRA) is a health questionnaire, used to provide individuals with an evaluation of their health risks and quality of life. Commonly a HRA incorporates three key elements – an extended questionnaire, a risk calculation or score, and some form of feedback i.e. face-to-face with a health advisor or an automatic online report.

The best known example of this proactive trend in medicine is probably the Framingham heart study. That groundbreaking effort has led to dozens of "risk factor" guidelines (Framingham Heart Study, http://www.framinghamheartstudy.org/). Its findings have led to significant changes in the practice of medicine through consideration of risk factors present in each patient, based on empirical data.

Familiar screening tests have been developed in the fields of breast cancer among women and prostate cancer among men. Other familiar procedures include simple screening indicators for diabetes, high blood pressure, and osteoporosis. Selection of screening tests vary depending on the person's age, sex, family history and other individual risk factors, such as the presence of obesity or the practice of smoking. There are hundreds of HRAs in use today for general health and for various specific conditions.

As an example of a recently developed broad spectrum screening procedure, in the field of aging, Hoogerduijn et al. (2007) reviewed screening instruments to identify risk

of functional decline among elders in hospital settings. The article discussed the history, validity, and reliability of clinical user-friendly instruments. Functional decline and loss of self-care skills is a common and serious problem in older hospitalized patients, resulting in change in quality of life and lifestyle.

> "Studies have shown that 30-60% of older people develop new dependencies in activities of daily living . . during their hospital placement and increased dependency of older people at home are the results. Not only are the personal costs high but also, in a rapidly growing older population, the impact on health care costs is also high." (p. 46).

Detection of such declines among individuals can lead to targeted intensified attention and stimulation to foster longer maintenance and expression of skills and independence.

Inouye et al. (2003) described a "burden of illness" score derived from a screening tool that took into account the cumulative impact of diseases, physiologic abnormalities, and functional impairments among elders.

Many screening and assessment tools have been developed in the general health field. However, most of the instruments and methods for health risk detection have been aimed at single conditions or groups of conditions. An example of this is given by Adriaanse et al. (2005), concerning screening for risk of Type 2 diabetes. This project focused on the psychological aspects of targeted screening for type 2 diabetes within the Hoorn Screening Study. While diabetes screening is now common, Adriaanse et al. were also concerned with the individual mental impacts of a "positive" finding in the initial screening.

Health risk assessment and screening instruments have usually been designed with the very desirable feature that the tools can be completed by professionals other than physicians. In this way, the non-MDs can act conceptually as "funnels," conserving precious resources by bringing only the most at-risk individuals into the full and costly scrutiny of the health care system. Simultaneously, they theoretically act to improve health by detecting conditions that would otherwise be missed – and the conditions would get worse. Both features should tend to produce superior health outcomes combined with fiscal conservatism.

Methods

Instruments

One of the earliest attempts to create a risk screening approach for people with developmental disabilities arose from efforts to analyze deaths among people moving out

of two large public institutions in Florida (O'Neall, Foster, & Butler, 1984). The people moved to seven smaller community based "cluster" facilities, and people moving to one of the seven, the "Hodges cluster," experienced an extraordinarily high mortality rate – approximately double that of the other clusters. State officials and scientists sought to find the cause(s) of this high death rate. The characteristics of the people as a possible explanation came up first. Perhaps the people who moved into the Hodges cluster had some special needs or conditions. Consultants created an instrument to examine this possibility. It was in essence a kind of HRA, or health risk assessment tool.

Using this new tool, the O'Neall study found that the individuals who moved to the Hodges cluster were not systematically different from the other people. In fact, the Hodges people tended to have somewhat fewer disabling conditions and health threats than the others. They found that the real cause of the deaths was inadequate health management practices at the Hodges cluster.

However, the study noted that the development of the tool itself was one of the most important outcomes of the work, because the tool when used as an HRA could be predictive of health deterioration or mortality. The study recommended establishment of a mechanism for early identification of risk factors for health destabilization.

The primary author of the survey used in the 1984 Hodges study continued to work on it. Over the course of further work in Florida, then Oklahoma and other states, the instrument was refined, revised, tested, and revised again. It was used on a large scale in Oklahoma, and on smaller scales in many other state and county systems. At that time, the Health Risk Screening Tool was a paper based survey. During the 00s, it was converted to a purely online version. According to Baker & Wilson (2007), the use of online technology can speed up and enhance the feedback loop, thus increasing the probability of averting health destabilization.

As of this writing, health risk screening approaches have been developed, tested, refined, improved, and applied over a 29 year period since the original Hodges study in Florida. The Health Risk Screening Tool, HRST, has been designed and applied for 21 years. It is the only tool of its kind in widespread use in the United States. It is simple and short, composed of 22 items. According to the HRST literature, the tool is aimed at prevention and early detection:

> The Health Risk Screening Tool (HRST) is a web-based rating instrument developed to screen for health risks associated with a wide variety of disabilities, including developmental disabilities, physical

disabilities, disabilities associated with aging, and many other conditions, which specifically affect systems of the body and the person's ability to engage in functional activities.

Its functions are listed as:

- Meets CMS health and safety requirements
- Detects health destabilization in vulnerable populations
- Prevents preventable deaths
- Quantifies level of health risk
- Determines range of clinical supports, services and surveillance needs
- Develops health baseline on individuals and groups

Because the HRST affects issues of life and death importance, it is crucial to know that it is accurate, i.e., that we can depend on the data to be genuine and trustworthy.

Reliability is the scientific term for accuracy. When we try to measure something about human life and health, are we doing it accurately? That is a very important question. Psychometricians (those concerned with measurement in health and human services) think in terms of three kinds of reliability.

1) Internal consistency reliability
   Do all the items in the instrument "hang together" – are they all correlated with one another – are they all getting at the same underlying dimension of reality that we want to measure?
2) Test-retest reliability
   Is the scale stable over time? (The items too.) A scale that varies all over the place from day to day isn't much good for anything.
3) Inter-rater reliability
   If two different interviewers (or data collectors or testers) collect data from one person, will they come back with similar data? If not, then we might as well stay home and flip coins.

In the study reported here, we have been able to test the first two forms of reliability. A test of the third is under way at the time of this writing, and a test of its predictive validity is in the design stage.

Procedures

HRS Inc., the corporate name of the instrument developers and users, now operates a database that covers more than 14,000 Georgia citizens with developmental

and intellectual disabilities. We requested and obtained three flat HRST data files from its work in Georgia, stripped of all identifiers to make it entirely HIPAA compliant. There were no restrictions on our methods of examining the data.

A randomized code was created for each individual by the HRST group, but this was never shared in any way with our research group. It enabled us only to organize the data over time. The data came as three files: March 2011, September 2011, and March 2012. At each time period, the HRS group pulled every person's complete 22 item HRST record and sent the raw data to us.

Once the data were received, the randomized individual codes were utilized to link the three files. We were then able to determine which people in the 14,000 had one observation, or two, or three – and to look at their scores over time.

The Georgia data were ideal for strong tests of the first two kinds of reliability. Internal consistency could be examined three times – once in March 2011, again in September 2011, and once again in Marcy 2012. The test-retest form of reliability could now be tested by comparing each person's HRST data over an 18 month period, to see if the data were stable over time.

Participants

The data set used for this study encompassed all the HRST "observations" for 14,000 people with developmental and intellectual disabilities receiving supports and services in the state of Georgia. Because we requested a "stripped" data file, there were no individual identifiers on the file, not even age, gender, or ethnicity. For this first examination, those indicators were completely unnecessary. Moreover, stripping them off the file avoided any and all concerns about individual privacy or confidentiality.

Hence we cannot describe the characteristics of the 14,000 participants in this part of our reliability investigations. We can say that the data file represents a population, not a sample, so considerations of sample validity do not apply. In future studies, we will work to include carefully limited individual characteristics – so that, for example, we can perform an analysis of the power of the HRST to "predict" or "explain" morbidity and mortality.

Results

Findings 1: Internal Consistency Reliability

For internal consistency, the most widely used test is Cronbach's Alpha. The higher the value of Alpha, the better the reliability. We calculated Alpha for all three sets of HRST data, and found:

Table 1: Internal Consistency of the Overall 22 Item HRST Scale

Data Set	Alpha
March 2011	.828
September 2011	.833
March 2012	.836

Alpha can only go up as high as 1.0, and the interpretation recommended by most psychometricians is:

Table 2: Cronbach's Alpha Levels of Acceptance [2]

Above 0.9	Excellent
Between 0.8 and 0.9	Good
Between 0.7 and 0.8	Acceptable
Between 0.6 and 0.7	Questionable
Between 0.5 and 0.6	Poor
Below 0.5	Unacceptable

Thus the Health Risk Screening Tool was found to have "Good" internal consistency reliability. This high level of internal consistency reliability was found at all three time periods. This strengthened to ability to infer that the finding was stable and valid.

Findings 2: Test-Retest Reliability for the Overall Scale

In the second reliability test, the data were compared over time – from March to September of 2011, and then from September 2011 to March 2012, to see if the scale

[2] Wikipedia entry 4/8/12 on Internal Consistency Reliability at http://en.wikipedia.org/wiki/Internal_consistency.

scores were stable over long periods of time. This was a very conservative test of test-retest reliability. Such tests are usually done over weeks, not months.

The data revealed that the HRST was very stable over time. The measure used for this test was simple correlation[3] – how similar the data were from one time to the next.

In the first analysis, the 22 items of the HRST were combined into a simple additive scale. Because each item is rated from 0 to 4 points,[4] the items could easily be combined without undue concern for weighting or scaling.

Table 3: Test-Retest Correlation Coefficients for the Overall 22 Item HRST Scale

Months Apart	Months Compared	Correlation
6 months apart	March 2011 to September 2011	$\mathbf{r_{1\text{-}2}} = .958$
6 months apart	September 2011 to March 2012	$\mathbf{r_{2\text{-}3}} = .962$
12 months apart	March 2011 to March 2012	$\mathbf{r_{1\text{-}3}} = .921$

This result showed that the HRST was very strong on test-retest reliability.[5] Correlations this high, over several months, are rare in social and health research.

Findings 3: Test-Retest Reliability, Item Analysis

The individual items in the 22 item HRST were also examined for stability over time.

Table 4: Test-Retest Reliability at the Item Level; HRST Item Correlations Over Time

[3] Pearson's product-moment correlation is the full name of the statistic.

[4] One item, Q, "Requirements for Licensing Intervention," was rated as a checklist with 7 boxes, yielding a 0 to 7 points. This departure was necessary for licensing considerations, and did not materially affect the nature of the overall scale.

[5] The way the HRST is administered is part of the reason for these extremely high correlations. As time goes on, changes are made to item scores, rather than the entire scale being re-administered. This procedure naturally minimizes "noise" or random fluctuations in item scores over time.

			Mar 2011 to Sep 2011	Sep 2011 to Mar 2012	Mar 2011 to Mar 2012
Risk Dimension		**Items**	r_{1-2}	r_{2-3}	r_{1-3}
Functional Status	A	Eating	0.961	0.966	0.929
	B	Ambulation	0.981	0.984	0.967
	C	Transfer	0.977	0.979	0.960
	D	Toileting	0.972	0.975	0.949
	E	Clinical Issues Affecting Daily Life	0.845	0.862	0.731
Behaviors	F	Self Abuse	0.924	0.925	0.853
	G	Aggression Toward Others and Property	0.913	0.918	0.838
	H	Use of Physical Restraints	0.906	0.876	0.787
	I	Use of Emergency Drugs	0.844	0.853	0.709
	J	Use of Psychotropic Medications	0.950	0.955	0.910
	K	Gastrointestinal Conditions	0.930	0.935	0.870
	L	Seizures	0.951	0.957	0.910
	M	Anticonvulsant Medication	0.955	0.952	0.910
	N	Skin Breakdown	0.912	0.925	0.840
	O	Bowel Function	0.937	0.939	0.879
Physiological	P	Nutrition	0.935	0.938	0.878
	Q	Requirements for Licensed Interventions	0.886	0.906	0.804
Safety	R	Injuries	0.780	0.805	0.600
	S	Falls	0.878	0.894	0.775
Frequency of Service	T	Professional Health Care Services	0.856	0.870	0.739
	U	Emergency Room Visits	0.780	0.786	0.586
	V	Hospital Admission	0.772	0.793	0.583
Average Correlations			0.902	0.909	0.819

Detailed inspection of this table reveals, first, that the overall correlations between items at six month intervals are higher than correlations at 12 months. The last row at the bottom of the table shows that item correlations between the first and the second observation, and between the second and the third, both six months apart, are nearly the same, at roughly .90. But between the first and the third observation, a year apart, the correlation is lower, at .819. This is expected and natural. Our study extended over a much longer period of time than is usually employed in test-retest studies, so over time, there must naturally be some genuine changes in the conditions of the people. These genuine changes tend to lower the correlations, and as time goes on, there are more changes, and the correlations must necessarily go downward.

The second indication supplied by Table 4 is that the correlations are generally quite high. Test-retest correlations, usually measured over much shorter time spans, are considered excellent above .90, and very good above .80.

The third issue of interest is the individual item reliabilities. Taking just the first column, from March 2011 to September 2011 ($\mathbf{r_{1\text{-}2)}}$, we see that the three highest reliabilities are for Ambulation (.981), Transfer (.977), and Toileting (.972) – three of the most easily observed of the 22 items – and among the least susceptible to change. In contrast, the three lowest are for Hospital Admissions (.772), Emergency Room Visits (.780), and Injuries (.780) – three of the items most susceptible to real changes in individual situations and events.

These data not only reveal strong reliability in general, they also suggest a pattern of real changes in actual life trajectories that correspond to lower test-retest correlation coefficients.

Discussion

This study supports to the inference that the HRST Online instrument and process display strong reliability of two kinds: internal consistency and test-retest. This is an important step toward establishing the accuracy and value of this health risk assessment for vulnerable people with developmental disabilities.

The usefulness of such a tool, designed to minimize preventable illness and death among vulnerable populations, could be high. Yet no such tool is in wide use in the nation; Georgia is one of the exceptions, adopting the HRST approach far ahead of other states. Because the HRST has been formally accepted as good practice by the Centers for Medicare & Medicaid Services, more states may adopt the HRST.

The studies done thus far are a significant step toward validation of the tool and the approach. However, this first study has significant limitations which must be made explicit, and there are more studies that need to be done.

Limitations

The test of internal consistency was appropriate and in fact was really three tests. This lent a kind of cross-validation to the reliability investigation. It was clear that the 22 items of the HRST "hang together" in that they are all closely intercorrelated.

However, our study of test-retest reliability was imperfect in two competing, but possibly important, ways. First, our test-retest method spanning 6 months was ultra conservative. Most studies of test-retest reliability are designed with days or weeks between observations, not months. When months pass, it is natural that real changes in people's condition will crop up, and this will tend to make the instrument appear less reliable. Actually, though, genuine changes in the health conditions of the people are not a function of the measurement instrument, but of the real world. This imperfection in our study design will tend to make the HRST look less reliable than it actually is.

On the other hand, the way the HRST is administered probably tends to inflate its test-retest reliability. The HRST Online is a fluid, dynamic process. Items are changed one at a time, at unpredictable intervals, whenever conditions change in individual lives.

This is not like the typical or ideal psychometric study of test-retest reliability. In the pure or ideal model, all 22 items would be administered at once, and then all 22 would be administered again, a few weeks later. Our situation with the HRST Online system is different. The administering nurses and case managers are asked to "update" the HRST at regular intervals, but they have access to the last round of data, and are asked only to change the items that have changed. This procedure surely tends to enhance test-retest reliability results beyond what would be expected from the classical designs.

One may argue that this is acceptable, even desirable, in the real world. Changing only the individual health characteristics that have changed is sensible and meaningful. One may also argue that the process takes into account errors in ratings at baseline; if the first time a person was rated, some of the ratings were inaccurate, later ratings should correct these and be reflected in the later corrections as unreliable data.

That being said, however, it remains important to be aware that this study of HRST test-retest reliability contained a feature that tended to inflate the correlation scores. This could not be overcome in the real data set with over 14,000 people. A small sample study of test-retest reliability might be contemplated in the future. However, we are acutely aware that such a study with 30, 50, or even 100 "new" ratings and re-ratings would suffer from other threats to validity, including sampling error and a departure from the "real world" method of HRST implementation.

A third limitation of our study is one of external validity. This study was done in only one state. Though the "sample" (actually a population) was large, there is no certainty that different state systems, with very different Medicaid and Waiver processes, would look similar to Georgia's. Moreover, with regard to the HRST instrument, Georgia has had more than five years to practice, learn, enhance, and refine the HRST process. In the first or second year of HRST in some new system or state, we cannot know that the reliabilities would be as high as those found here. Hence, we would urge new implementers to include reliability studies as a required and ongoing feature from the start.

A fourth limitation arises from the fact that the data file for this study did not include any demographic information. For all we know, the tool may work differently for men and women, or young and old, or among different races. We have no tests of these possibilities in this limited study. This does not change the fact that overall high

reliabilities were found for more than 14,000 people. Nevertheless, future analyses could and should include such data so that such possibilities can be explored.

Further Research Needed

The next natural step in examining the reliability of the HRST Online is to conduct a rigorous study of interrater reliability. That third kind of reliability was not included in this study because the structure and process of the HRST Online project did not permit it. There are no instances of "two raters acting independently" within Georgia's 14,000 person system, as far as we could find. A new and separately designed study will have to be conducted.[6]

Following that test of the third form of reliability, a study of validity should be initiated – not just as a one time analysis, but as an ongoing practice. We recommend that the first validity study be aimed at mortality, asking the question "Do any of the HRST items, or the total scale, predict the probability of mortality among the 14,000 Georgia citizens in the database?" This kind of study would begin show the degree to which the HRST can accurately raise "red flags" of warning for people with a certain characteristic or combination of characteristics.

[6] Such a study has already been designed, a study site determined, and specific work on participant selection has begun. It will, however, require several months, at minimum, to complete.

References

Adriaanse, M., Dekker, J., Spijkerman, A., Twisk, J., Nijpels, G., van der Ploeg, H., Heine, R., and Snoek, F. (2005). Diabetes-Related Symptoms and Negative Mood in Participants of a Targeted Population-Screening Program for Type 2 Diabetes: The Hoorn Screening Study. *Quality of Life Research , Vol. 14, No. 6 (Aug., 2005)*, pp. 1501-1509

Alexander G (2000). Health risk appraisal. *The International Electronic Journal of Health Education 3 (Special):* 133–137.

Baker K, DeJoy D, and Wilson M. (2007). Using online health risk assessments. *The Journal of Employee Assistance.* April 2007.

Bassuk, E.L., & Gerson, S. (1978). Deinstitutionalization and mental health services. *Scientific American, 238*, 46-53.

Conroy, J. (December, 1993). *Statement of the Findings of James W. Conroy, Ph.D., from the Pennsylvania Office of Mental Retardation's 1992 Monitoring Data*. Expert testimony presented to Judge Raymond Broderick in the District Court of the Eastern District of Pennsylvania, bearing on contempt proceedings against the City of Philadelphia, as part of Halderman v. Pennhurst.

Conroy, J., & Adler, M. (1998). Mortality Among Pennhurst Class Members, 1978 to 1989: A Brief Report. *Mental Retardation, 36, 5*, 380-385.

Conroy, J. W., & Bradley, V. J. (1985). *The Pennhurst Longitudinal Study: A report of 5 years of research and analysis.* Pennhurst Study Report PC-85-1. Philadelphia, PA: Temple University Developmental Disabilities Center/UAP. Boston: Human Services Research Institute.

Framingham Heart Study website, http://www.framinghamheartstudy.org/ accessed 8/20/12),

Halderman v. Pennhurst State Sch. & Hosp., 446 F.Supp. 1295 (E.D. Pa. 1977).

Hoogerduijn, J.G.. Schuurmans, M.J., Duijnstee, M.S., DeRooij, S.E., and Grypdonck, M.F. (2007), A systematic review of predictors and screening instruments to identify older hospitalized patients at risk for functional decline. *Journal of Clinical Nursing, 16:* 46-57. Doi 10.1111/j. 1365-2702.2006.0159.x.

Sharon K. Inouye, Sidney T. Bogardus, Jr., Gail Vitagliano, Mayur M. Desai, Christianna S. Williams, Jacqueline N. Grady and Jeanne D. Scinto (2003). Burden of Illness Score for Elderly Persons: Risk Adjustment Incorporating the Cumulative Impact of Diseases, Physiologic Abnormalities, and Functional Impairments**.** *Medical Care, Vol. 41, No. 1* (Jan., 2003), pp. 70-83.

Lakin, K.C., Larson, S.A., Salmi, P., and A. Webster. (2010). *Residential Services for Persons with Developmental Disabilities: Status and trends through 2009*. Minneapolis: University of Minnesota, Research and Training Center on Community Living, Institute on Community Integration.

Mary Ann Lewis, Charles E. Lewis, Barbara Leake, Bryan H. King, and Robert Lindemann (2002). The quality of health care for adults with developmental disabilities. *Public Health Rep. 2002 Mar-Apr; 117(2)*: 174–184.

Nowell, N., Baker, D., & Conroy, J. (1989). *The Provision of Community Medical Care in Philadelphia and Northeastern Pennsylvania for People Who Live in Community Living Arrangements and with Their Families*. Report prepared under a grant from the Pew Charitable Trusts. Philadelphia, PA: Northwest Center.

O'Neall, Linda R., R.E Foster, and M.B. Butler (1984). *The Hodges Report*. Brehon Institute for Human Services.

University of Minnesota (2011). *Behavioral Outcomes of Deinstitutionalization for People with Intellectual and/or Developmental Disabilities: Third Decennial Review of U.S. Studies, 1977-2010*. Policy Research Brief Volume 21, Number 2. Research and Training Center on Community Living, Institute on Community Integration College of Education and Human Development, University of Minnesota.

William F Sullivan, John Heng, Donna Cameron, Yona Lunsky, Tom Cheetham, Brian Hennen, Elspeth A Bradley, Joseph M Berg, Marika Korossy, Cynthia Forster-Gibson, Maria Gitta, Chrissoula Stavrakaki, Bruce McCreary and Irene Swift (2006). Consensus guidelines for primary health care of adults with developmental disabilities. *Canadian Family Physician November 2006 vol. 52 no*. 11 1410-1418.`