TESTING THE RELIABILITY AND VALIDITY OF THE WISCONSIN LONG TERM CARE FUNCTIONAL SCREEN.

Wisconsin Department of Health and Family Services
Office of Strategic Finance

Center for Delivery Systems Development
One South Pinckney Street, Suite 340
Madison, WI 53703

TESTING THE RELIABILITY AND VALIDITY OF THE WISCONSIN LONG TERM CARE FUNCTIONAL SCREEN.

A Paper prepared as part of Wisconsin's application to the Health Care Financing Administration for approval of the Long Term Care Functional Screen for use by Wisconsin's Home and Community Based Waivers, in Family Care demonstration counties.

ABSTRACT

Two of the leading measures of an instrument's quality are reliability (consistency) and validity (the degree to which it measures the phenomenon it aims to measure). In an effort to document the quality of the Wisconsin Long Term Care Functional Screen (LTCFS), two studies to measure the reliability and validity of this centerpiece of Family Care were undertaken during 1999.

The reliability study utilized an inter-rater reliability design and consisted of a pilot phase and a main study. During the pilot phase, dual screens were administered to a sample of 98 clients in 9 Resource Centers in Wisconsin. The two screen were administered by different workers, independently and within a short time frame, to each client in the sample. The analysis focused on the level of agreement between the two screeners on core items of the LTCFS that are used in the process of establishing Family Care eligibility. The results of the analysis and the following explanations from the Resource Centers' workers, had led to the modification of the LTCFS and the development of its current Version 2.

The main Inter Rater Reliability Test (IRRT) study, had used a similar design and dual screening were conducted on a different sample of 96 clients. The sample included cases from all 3 Family Care target groups. The statistical analysis focused on 21 core items in the areas of ADL, IADL, Cognition, and Communication, and utilized two primary Statistics: the Chi-Square test of association, and Cohen's Kappa correlation coefficient (for agreement between two raters). The analysis established highly significant levels of association and correlation. The Chi-Square results were all significant at the $p<0.0001$ level and the magnitudes of the Cohen's Kappa which ranged from 0.59 to 0.92, were also statistically significant given the sample size and Degrees of Freedom.

The results of the reliability study suggest that the LTCFS is a consistent tool that is expected to yield similar results when administered to clients in similar circumstances, within a short time frame.

The methodology of building the LTCFS, had contributed to its second measure of quality, validity. The involvement of in-house and external experts, the borrowing of tested components from established instruments, and the consulting of written materials from other states and independent research sources, have greatly enhanced the face, content, and construct validity of the tool. Nevertheless, the Center undertook a special

study that utilized a criterion validation approach, considered by many to be a "harder" measure of validity.

The criterion that was selected for the validity study was the BQA Nursing Home Level of Care (LOC). The BQA LOC has been accepted by HCFA and, in fact, is a dual track process, one for frail elders and people with physical disabilities and the other for clients with developmental disabilities. The purpose of the study was to measure the agreement between the LOC obtained by using the LTCFS logic, and the LOC obtained by BQA (the criterion). Like the reliability study, this study also consisted of a preliminary, pilot phase, and a main study.

In the preliminary study, two samples of 151 nursing home cases and 131 DD cases for which BQA LOC were established, and who also completed a LTCFS, were selected. A research assistant, who was unaware of the BQA LOC, used extracts of information from the LTCFS and the Screen's logic to establish a LOC for the same clients. Finally an analysis was conducted to determine whether all clients who had attained a BQA LOC also attained any LOC on the LTCFS. The results of this "in-out" study (84% agreement in NH cases and near perfect agreement for DD cases), were used to improve the LTCFS LOC logic.

In the main study, two different samples of 79 nursing home and 67 DD clients for whom BQA LOC was established during 1999, were selected. The cases were selected to represent all regions of the state as well as institutionalized and community based clients. Four nurses reviewed the BQA documentation for these clients and converted core information in these records to LTCFS information format. A Center worker, who was unaware of the BQA LOC determinations then ran these information extracts through the LTCFS logic to assign these clients an independent LOC.

The two Statistics that were used to measure the agreement between the LTCFS determinations and those made by BQA were the Chi-Square test of association, and the Gamma-Kruskal correlation coefficient (for ordinal data). The analysis yielded significant levels of association between the LOC attained by BQA and the LTCFS, as measured by the Chi-square test. The Gamma coefficients for both the nursing homes, and DD samples were 0.93, and given the samples sizes (converted to Degrees of Freedom) are highly significant.

The results of the validity study suggest, that the LTCFS can be used in lieu of the BQA process to establish nursing home levels of care for both frail elder and people with physical disabilities, as well as the developmentally disabled population.

Introduction

The Wisconsin Long Term Care Functional Screen (LTCFC) is a multi-purpose center-piece of the Family Care demonstration project. Family Care is an innovative and

integrated service delivery system designed to meet long term needs of three populations at risk: Persons with physical disability, persons with developmental disability, and frail old adults. The program aims to consolidate existing and fragmented programs, merge separate funding streams, and enhance consumer choice. Family Care is a cooperative effort drawing on the support of Federal, State, and County governments, and on the involvement of advocacy and consumer groups throughout the state. In its current phase, Family Care is a demonstration project that is planned for nine Wisconsin counties. There are, however, plans to expand the program to additional areas of the state, should the current demonstration prove its effectiveness and efficiency.

The Long Term Care Functional Screen: purpose and objectives

From the onset, project planners have sought to develop a multi-purpose instrument that would meet a variety of diagnostic, research, and programmatic needs that are associated with operating and managing Family Care. This instrument is the Long Term Care Functional Screen (LTCFS).

The LTCFS is one of several screening instruments currently used in Wisconsin for identifying long-term care needs and program eligibility. Most of the other instruments were designed to support placement decisions related to specific settings (typically nursing homes or the community), or are oriented toward assessing the needs of specific populations (Developmentally Disabled, Frail Elderly, etc.). The new LTCFS is seen as a more integrated tool, intended to support a broad array of services, accomplish multiple goals, and serve multiple risk groups, along the entire continuum of long term care settings. Furthermore, the LTCFS aims to reduce the need for multiple, often obtrusive, contacts with prospective clients, through the streamlining of questions to clients, and decision logic.

In its design, the LTCFS aims to measure and integrate a client's functional and medical aspects. The screen is made up of separate modules that record wide-ranging data on a client. Data cover the areas of demographics, living arrangements, ADL and IADL functioning, diagnosis, medical tasks, communications and cognition, behavioral attributes, and risk. The data in each module can be studied alone to gain insight into one aspect of a client's needs and circumstances. The various decision logic that are associated with the screen, including the logic to determine Family Care eligibility level, and Nursing Home level of care, link data from several modules in hierarchical processes.

The LTCFS will be offered to all Resource Center contacts with a prospective long-term care need. It will also be required for all prospective admissions to a Nursing Home, DD facility, or a CBRF as part of the Pre Admission Counseling (PAC) process. Should a prospective client, who has completed the LTCFS, select a Care Management Organization (CMO) as his/her long term care arrangement, the information gathered by the LTCFS will be used as part of a comprehensive client assessment and care planning that will be performed for all new clients of the CMO, further reducing redundancy, obtrusion, and administrative cost.

The objectives for administering (and gathering data on) the LTCFS include:

- Determining functional eligibility for Family Care
- Evaluating potential clients' functional status
- Determining potential clients' living arrangement preferences
- Assigning clients to the appropriate Nursing Home level of care
- Upon the completion of planned research, assigning clients to appropriate cost bands to allow for accurate and adequate capitated payments to the responsible agency and its providers
- Collecting base line and on-going data for the evaluation of client outcomes and assessment of the program's cost effectiveness.
- Monitoring and enhancing of program quality.
- Detection and early referral of special and urgent needs like adult protective services.
- Establishing a critical consultation "gate keeping" point for clients and family members that are contemplating transitions, especially to institutional settings
- Collecting data on clients' financial resources, to assist in LTC choice counseling pre-admission and other benefit consultations.

Screen development and history

The current version of the LTCFS is the product of more than two years of collaborative efforts led by the Center for Delivery Systems Development in the Wisconsin Department of Health and Family services. The team working on the development of the Screen combined multi-disciplinary and multi-agency perspectives and expertise. At different points during the process the effort involved nurses, social workers, gerontologists, developmental psychologists, information specialists, and other professionals representing social service agencies, nursing homes, community health, and other human service organizations. Much of the developmental work was conducted with the participation of clinical experts and experienced workers from the field.

Although the screen is a new and innovative instrument, its development drew heavily from some of the leading screening instruments currently in use. Many items were designed for compatibility with items from current instruments, to enhance construct validity and to increase compatibility with other long-term programs. Among the instruments used by Wisconsin, that were consulted during the construction of the

Screen, are the Minimum Data Set (MDS), the Outcome Assessment Information Set (OASIS), the Wisconsin COP (Community Options program) Functional Screen, and the Wisconsin Medicaid Home Care Assessment. In addition, screening instruments from seven other states (Oregon, Alabama, Illinois, Vermont, Massachusetts, Washington, and Maine), and several federal agencies and publications were consulted.

To further enhance (content) validity, earlier versions of the LTCFS were submitted to panels of external experts for review. Overall, more than 100 experts from outside the Center for Delivery Systems Development reviewed all or parts of the screen and offered modifications and other feedback.

Although the face and content validity of the LTCFS have been documented throughout the process, Family Care managers as well as HCFA project officers have sought to obtain harder, more empirical evidence of the quality of the LTCFS and the screening process. This in advance of approving the use of the Screen, in lieu of current instruments, for a range of clinical, evaluative and fiscal purposes.

The need to document quality

Two of the most acceptable criteria of scientific rigor for any instrument used in diagnosis and research, are reliability and validity. Simply defined, reliability refers to the consistency of the measuring instrument and its ability to deliver the same measurements under similar conditions. Validity refers to the accuracy of an instrument and its ability to truly measure the phenomena that it purports to measure. An instrument can be reliable without being valid, but all valid instruments are by definition reliable. Nevertheless, it is customary, when examining the use of a new instrument, to study reliability and validity independently.

Although there is wide agreement on the basic definitions of reliability and validity, there are different approaches to their measurement. The following sections address the issues that are involved in the testing of the LTCFS reliability and validity, and report the findings from the extensive research that Wisconsin has performed to test each.

RELIABILITY: METHODOLOGY AND DESIGN

The procedure that was selected by Wisconsin to evaluate the reliability of the LTCFS was the Inter-Rater Reliability Test (IRRT). This procedure involves the administration of an instrument to the same subject by two different observers/raters, within the same time frame. The aim is to measure the level of agreement between the two raters on key components of the instrument. The stronger the agreement between the raters, the more confident one can be that the instrument is reliable (consistent). The IRRT has several advantages including the ability to engage the two screeners in the follow up process of studying the causes for disagreements and improving the instrument.

By administering the same instrument, within the same time frame, we attempt to reduce the potential impact of maturation effects (Stanley and Campbell, 1967), e.g., actual and

time related changes in client condition. The IRRT assumes that both raters are adequately trained. When this condition is not met, any disagreement that is detected may be attributable to the efficacy of the raters rather than the reliability of the instrument. To ensure rater efficacy, extensive training was provided to all participating Resource Centers' workers who engage in the performance of the LTCFS. The training was provided on the earlier version of the Screen, before the pilot phase of the study, and again, on the current version, before the main study.

A word of caution relates to the possibility that both raters can agree (provide the same score) but both be wrong by differing from the true answer. Such a problem can often be explained by lack of validity (and is addressed in the sections that cover validity).

The implementation of the IRRT by Wisconsin consisted of two phases: the pilot phase, and the main study:

The pilot phase was carried out using the earlier version of the LTCFS and was based on a final sample of 103 cases. The cases were selected from all 9 original Resource Center counties and represented all three Family Care Target Groups (frail elders, people with physical disabilities, and people with developmental disabilities). A total of 36 workers participated as raters in the study and each case was reviewed independently by two raters in the same county.

The analysis consisted of the statistical measurement of agreement between the raters on the core items in the LTCFS, which are used in the determination of functional levels. The statistical procedures are described in detail as part of the main study. Following the statistical analysis, clinical staff from the Center reviewed each pair of screens and identified actual discrepancies (disagreements) between the screeners. The screens containing the discrepancies were sent back to the Resource Centers and the screeners were asked to explain their source. The explanations, together with the results of the statistical analysis were used to modify several flawed items and to develop the current version of the LTCFS.

For the main IRRT study, a different sample of 96 cases/subjects was randomly selected. Subjects were either new Resource Center contacts, or existing clients that were due for review. Screens for the 96 subjects were also conducted in the original 9 Resource Center pilot counties (Milwaukee, Fund du Lac, Kenosha, La Crosse, Marathon, Portage, Richland, Trempealeu, and Jackson), and the cases represent all three Family Care Target Groups. Because the test focuses on the internal dynamics and interactions within the instrument, the impact of external variables (such as the agency size) on the establishment of internal reliability is small, hence the relative unimportance of each county proportion in the sample. More important are the sample's sufficient size, its random selection, the inclusion of members from all Family Care Target Groups in the test (Cochran, 1977), and the relative lack of familiarity of the raters with the subjects to reduce bias. All efforts were made to meet these conditions. Of the 192 screens (two screens for each subject), 65% were administered to frail elderly contacts, 21% to people with developmentally disabilities, and 14% to people with physical disabilities.

Because the IRRT aims at measuring levels of agreement within the same time frame (Campbell & Stanley, 1967), the administration of the two screens, for each person, occurred within 2 weeks of each other.

Two statistical tests were used to measure agreement between the pairs of raters. The Chi-Square test is used to determine whether the association between two variables (the scores of the two raters), is due to chance. This determination is based on the calculated value of the Chi-Square and on the number of degrees of freedom, which, in turn, are determined by the number of cells into which observations can potentially fall.

The second test, Cohen's Kappa (Cohen, 1997; Howell, 1992; Schuerman, Rzepnicki, & Littell, 1994) measures the level of correlation between the scores of two raters, e.g., the ability to predict the score of one rater on a certain measure, by knowing the score of the other. The Kappa coefficient has been amply documented and used extensively to measure agreement between two raters in a variety of fields. Examples include the measurement of agreement between husbands and wives on shared experiences, between employers and employees on a desired course of action, and between physicians on patient diagnosis (Altman, 1991). The Kappa's procedure is included in both SAS (Statistical Analysis System) and SPSS (Statistical Package for the Social Sciences) computerized statistical packages.

The 21 items that were selected for analysis, because of their centrality to the eligibility determination process and to service planning, all fall into ordinal definitions (lower scores imply more independence or lower severity, and higher scores imply more dependence or higher severity), where weighted measures are more appropriate. In the context of the Kappa coefficient, weighting implies that smaller differences between raters have a smaller negative impact on the magnitude of the correlation, than larger ones. Thus, for example, a difference of one point between the raters will have less of an effect on the coefficient, than a difference of two points.

Although the Kappa benchmarks for acceptance depend on the question under study, and on the cost of making erroneous decisions, Landis and Koch (1977) suggest the following coefficient benchmarks: poor (<0), slight (0-0.19), fair (0.20-0.39), moderate (0.40-0.59), substantial (0.60-0.79), and near perfect (0.80-1.00). As in the case of other measures of correlation, a separate test was performed to assure the significance of each coefficient.

The computerized statistical package that was used to calculate the Kappa coefficient, Analyse-It, is a comprehensive statistical package that was developed primarily for Excel spread sheets. The program automatically excludes from the analysis items with missing data. Thus, when data from one or both raters was missing for any given item, the case was excluded (only) from the analysis of that item.

Findings

Calculations for Cohen Kappa coefficients were performed on (21) items that are included in the clinical modules of the LTCFS. The following tables summarize the findings for these core items:

1) ADL/IADL

ITEM DESCRIPTION	N	KAPPA*	CHI SQUARE	CHI SQUARE DF	p Value
Bathing	96	0.79	93.26	4	<0.0001
Dressing	96	0.79	100.14	4	<0.0001
Eating	96	0.68	82.12	4	<0.0001
Toileting	96	0.76	70.55	4	<0.0001
Mobility	96	0.69	88.52	4	<0.0001
Transferring	96	0.78	132.19	4	<0.0001
Incontinence, bladder	96	0.68	0.44	1	<0.5 NS
Meal preparation	96	0.74	129.75	4	<0.0001
Money management	96	0.92	110.97	1	<0.0001
Telephone use	96	0.73	48.2	1	<0.0001

2) TRANSPORTATION AND EMPLOYMENT

ITEM DESCRIPTION	N	KAPPA*	CHI SQUARE	DF	p Value
Transportation	96	0.65	99.80	4	<0.0001
Employment	96	0.92	324.00	16	>0.0001

3) COGNITION AND COMMUNICATION

ITEM DESCRIPTION	N	KAPPA*	CHI SQUARE	DF	p Value
Communication	96	0.83			
Resistance to care	96	0.59	24.67	1	<0.0001
Long term memory	96	0.82	135.40	2	<0.0001
Short term memory	96	0.82	60.48	1	<0.0001

4) BEHAVIORS/SYMPTOMS

ITEM DESCRIPTION	N	KAPPA*	CHI SQUARE	DF	p Value
Wandering	96	0.61	62.73	4	<0.0001
Self injurious	96	0.65	167.49	9	<0.0001
Violent/Offensive	96	0.74	147.55	9	<0.0001
Mental Health	96	0.70	77.74	9	<0.0001
Substance abuse	96	1.00	65.81	1	<0.0001

*All 21 Kappa coefficients in the 4 tables are significant at the $p<0.01$(DF = 94)

Interpretation and discussion of findings from the reliability study

Using the Landis-Koch guidelines as benchmarks, all items, with the exception of Resistance to Care met the "Substantial" threshold of reliability and several had exceeded this standard. In addition, all coefficients are statistically significant. With the exception of one item (Incontinence, Bladder), all the items on our list also show highly significant associations between the scores of the two raters, as measured by the Chi-Square test.

The general interpretation of the weighted Kappa coefficients is quite simple: the fewer the number of disagreements between raters for each item, and the smaller the magnitude of the discrepancies (when disagreements are present), the higher the value of the Kappa coefficient. In addition, the way potential outcomes (potential outcomes for an item are the range of all types of agreements and disagreements) are distributed in the contingency table, which is the basis for Kappa calculation, can also affect the coefficient. Our additional analysis (calculating the percentages of disagreements for each of the 21 items), suggest that the magnitude of many Kappa measures that we obtained, tend to be somewhat conservative and actually underestimate the levels of agreement. The results of that analysis are not shown here, but documentation is being kept.

The 21 items that are included in the four tables are not the only items that make up the LTCFS. They are, however, the only items that display sufficient score variation to allow meaningful statistical analysis. A case in point is the LTCFS modules that gather data on diagnosis and medical needs. Although the sample size for the reliability study meets scientific conditions for performing the required analysis, the prevalence of most diagnosis and medical needs is too small and sketchy to enable sufficient variation. Under such conditions, the validity of most statistical analysis is highly questionable, and the ability to infer from the sample to the general population is fraught with danger.

To assure the continuing quality of the LTCFS, the Center intends to monitor the reliability through continuous training and technical support, as well as small-scale studies. We have prepared several questionnaires that already have been sent to the sites

where research was conducted. Raters are asked to review the discrepancies and select explanations from the following list:

A. A mistake by one of the screeners
B. A misunderstanding by one of the screeners
C. Unclear wording on the screen
D. Inadequate training on the screen
E. Consumer and family members provide different information
F. Consumer's condition is unstable and screeners made different interpretations.
G. Consumer's condition actually changed between the two administrations of the screen

The Center plans to continue the analysis of discrepancies and explanations, in order to improve training, technical assistance, and manual writing.

VALIDITY: DEFINITION AND APPROACHES

The general definition of validity refers to the ability of an instrument to truly measure that which it purports to measure. Validity is a basic requirement that every scientific instrument or process must meet before its use is accepted. Commonly, in selecting an instrument researchers and practitioners can exercise one of three options: they can use an instrument, the validity of which is already recognized, they can adapt an existing instrument, or they can develop their own. When selecting the third option, establishment of the instrument's validity (and reliability) must be documented before it can be accepted. Although the LTCFS has borrowed from several existing instruments, it is fundamentally an original tool and must therefore be validated.

There are four common approaches to defining and measuring an instrument's validity:

- Face validity is the extent to which a measure is subjectively viewed by experts as representing a concept. The judgment usually involves more than one expert. When there is lack of consensus among the experts regarding a certain item or a scale, the item in question is usually discarded or modified.

- Content validity is somewhat similar to face validity and is also based on expert opinion. Like face validity content validity also addresses the issue of concept representation by the instrument, and in addition focuses on the level of instrument completeness. This approach seeks to ensure that a scale or another instrument contains all the relevant items that are needed to measure the phenomenon.

- Construct validity seeks to link a person's scores on the instrument under development to his/her scores on other indicators that are known to be associated with the phenomenon in question. For example, in developing a scale to measure depression among the elderly, we expect that the (depression) scores obtained by using the new scale will correlate with independent measures such as informal support and social integration for the same people.

- Criterion validity relates to the instrument's ability to agree with some criterion external to it. The criterion is, in effect, an alternative measure of the same phenomenon, which has already been accepted as valid. The criterion (in the context of research, the dependent variable) can exist in the present (concurrent criterion validity) or materialize in the future (predictive validity). An example of the latter would be an attempt to validate a job-interviewing instrument by looking at latter job performance.

As described earlier, the team that developed the LTCFS had applied the first three approaches as integral parts of its work. Thus, throughout the process, the team obtained expert opinion that the items on the screen indeed measure client conditions and needs that are essential for the establishment of levels of care. Face validity was also achieved by "crosswalks" (visual comparisons of the items in the screen with items in other accepted instruments). The Center keeps documentation related to the steps taken to assure content validation of the Screen.

In the context of the LTCFS, the matter of the instrument's completeness (content validity) was another focal point during the development phases. Outside experts were asked to determine whether the list of items in the screen (those relating to ADLs, IADLs, Cognition, Communication, Behaviors, Diagnosis, and others), are exhaustive, inclusive, and sufficient to accurately establish a level of care. The documentation of this content validation process is being kept by the Center and can be referenced as needed.

The underlying structure of the LTCFS is based on construct validity. Most items in the instrument were included because they are established measures of functional and medical needs, and already correlate with the resources (such as staff time) that are required to meet those needs. Furthermore, an analysis of the LTCFS logic to determine nursing home level of care shows that such logic is anchored in both social and nursing research.

VALIDITY: METHODOLOGY AND DESIGN

The criterion

The previous section described how the LTCFS already meets the basic conditions for face, content, and construct validity. To further enhance the validation process, Wisconsin conducted a more rigorous study to test the LTCFS criterion validity. In the case of Family Care, the concurrent criterion validity approach is seen as the most practical and cost effective among the "harder", more rigorous approaches to validation. The actual criterion that was selected for the study is the Wisconsin's Bureau of Quality Assurance's (BQA), Nursing Home Level of Care determination process. Although not an instrument in itself, BQA's NH Level of Care (LOC) is the outcome/product of an accepted and well-documented process that utilizes a structured instrument: the MDS (Minimum Data Set), as well as extensive physician narratives. The BQA LOC is population specific and the determination process for frail elders and people with

physical disabilities differs somewhat from the process for people with developmental disabilities. The study, therefore, aims to validate the LTCFS for use with members of both populations.

BQA's Level of Care is an ordinal scale with three levels of care for the elderly and physically disabled, and three levels for the developmentally disabled.

The following levels apply to frail elders and people with physical disabilities (PD):

1) Intermediate Care Facility (ICF). Made up of the two sub levels of ICF-1 and ICF-2. ICF is the lowest NH Level of Care.

2) Skilled Nursing Facility (SNF)

3) Intensive Skilled Nursing (ISN). This is the highest level.

For people with developmental disabilities (DD), BQA sets the following three severity levels:

1) DD-3 (the lowest level)

2) DD-2

3) DD-1. Made up of DD1-A for the medically needy and DD1-B (same level of severity, but for the behaviorally challenged). This is the highest level.

The LTCFS, unlike the BQA process, is a stand alone structured instrument but its range of potential LOC outcomes is identical to that of the BQA process. Our main study aims to validate the LTCFS by comparing its outcomes with those of the BQA process, for the same clients, for the two populations (frail elders and people with PD, and people with DD). Because specific items in the two processes do not correspond to each other directly, the validation of single items, by using this methodology, is not possible.

<u>The study's two phases: Description</u>

Like the Inter Rater Reliability Test, the LTCFS-BQA criterion validation study was conducted in two phases:

The LTCFS-BQA pilot phase examined whether a client who met any Nursing Home (or DD) Level of Care set by the BQA process, also met any NH (or DD) Level of Care on the LTCFS. As part of the analysis 151 nursing home cases and 131 DD cases for which both BQA and LTCFS Level of Care were established, were selected. Findings were that of the 151 nursing home clients who attained one of the three BQA Levels of Care, 127 (84%) also attained one of these levels on the LTCFS. Of the 131 DD cases that had one of the three BQA DD Level of Care, all (100%) also attained one of these levels on the LTCFS. Like the findings from the first phase of the IRRT, results from the first phase of

the validation study were used to correct and improve the logic to determine Level of Care in the screen. In addition, the findings were used to correct the design of the second phase of the validation study. An example of such correction is the decision to neutralize the effect of time in the design of the main study. This correction was necessitated by the finding that many of the disagreements among the nursing home population were associated with significant time differences between BQA's determination and the administration of the LTCFS.

The LTCFS-BQA main study aimed to compare specific levels of care in BQA to those attained by the same clients on the LTCFS. For this analysis two samples of clients, one containing 79 frail elders and people with PD, and the second containing 67 people with DD, were selected at random from the universe of clients for whom a level of care was established by BQA during 1999.

To reduce the impact of time differences between the BQA and LTCFS processes, on determinations (one of the primary lessons from the preliminary study), the main study utilized a time neutral approach, which consisted of the following steps:

1) Four nurses from BQA, who specializes in LOC determinations, studied the BQA records of the clients in the sample.
2) The nurses converted those elements in the BQA record that were used to set BQA LOC, into LTCFS format, in a manner that would enable a person who is proficient in the use of the LTCFS to generate a Level of Care based on the LTCFS logic.
3) A clinical expert from the Center ran the extracted information through the LTCFS logic to independently establish a LTCFS LOC for the case. The LTCFS expert was unaware of the BQA LOC that was assigned to the case.
4) The BQA and LTCFS LOC determinations were compared and tested for agreement.

The samples

Two samples, one from frail elders and PD population, and one from the DD group, were selected at random from the universe of all clients for whom a BQA Nursing Home or DD LOC were established during 1999. The sampling methodology utilized a stratified approach to ensure that cases with all BQA levels of care were represented. Stratification procedures were warranted since LOC distributions are skewed. An example is the distribution of aged and physically disabled clients, among the three levels included in the study. In 1997, for example, nearly 85% of all nursing home clients attained an SNF level of care, while only 0.5% were assigned the level of ICF. In order to demonstrate the validity of the LTCFS across levels of care, the nursing home sample was stratified to ensure sufficient representation of clients from all levels. To demonstrate validity across service settings the sample included, in addition to nursing home clients, clients in community based programs such as Partnership for whom BQA LOC was established. The final sample for the aged and physically disabled thus included 79 cases (25 at the ICF level, 29 SNF, and 25 ISN).

The DD sample included 67 clients (34 from the DD1, the most common level, 18 DD2, and 12 DD3, the least common category). The smaller DD sample reflects the smaller universe of DD BQA determinations.

In setting both sample sizes, primary consideration given to meeting the minimum condition for the performance of the Chi- Square test of association. The test requires a minimum of 5 cases per cell, and with two tables of 3X3 (3 BQA levels against 3 LTCFS levels) our samples both exceed the Chi- Square minimum of 45.

Statistical analysis

The statistical comparisons of the two outcomes utilized non-parametric procedures designed for the analysis of ordinal data. Statistical procedures include the Chi Square measure of association (which was also used in the IRRT study), and the Goodman and Kruskal's Gamma coefficient, for measuring the correlation between the BQA and LTCFS determinations. The Gamma test was determined to be the appropriate choice for the correlation of ordinal observations (the concept of level of care, by definition, implies ordinality). Although the establishment of acceptability levels (benchmarks) for the Gamma, depends on an array of factors including sample size, levels of 0.7 and above are regarded as acceptable.

Findings

1) LEVELS OF AGREEMENT BETWEEN THE OUTCOMES OF THE BQA LOC DTERMINATION PROCESS, AND THOSE OF THE LTCFS

LOC TYPE	N	GAMMA	CHI SQUARE	CHI SQUARE DF	P Value
NURSING HOME	79	0.938*	58.35	4	<0.0001
DD	67	0.938**	523.30	4	<0.0000

*Gamma DF=77, $p<0.01$
** Gamma DF=65, $p<0.01$

Interpretation and discussion of findings from the validity study

The second phase of the criterion validation study utilized a time neutral modified design, and yielded highly significant correlation coefficients between the outcome of the BQA determination process and that of the LTCFS. The Chi-Square test of association also yielded highly significant levels of association between BQA LOC and the LOC obtained by using the LTCFS logic. These findings apply to both analysis groups (frail elders and people with PD, and people with DD). The 0.938 Gamma coefficients that were obtained for both samples, translate into an R^2 factor of 0.88. The R^2 identifies the proportion of the variation in the scores of the BQA raters that can be explained by the variation in the

scores of the LTCFS determinations. In more practical terms, the higher the factor, the better is our ability to predict the LTCFS LOC when the BQA LOC is known. A high factor, such as the study had found, also implies strong underlying similarities between the two processes (LTCFS and BQA).

SUMMARY AND CONCLUSIONS

The approach employed by the Center: to conduct each (reliability and validity) study in two phases, aimed to ensure a more rigorous design as well as continuous review and improvement of the LTCFS through analysis and feedback. Although both reliability and validity studies yielded significant results, it is the Center's intent to continue to review the instrument on an ongoing basis, and to assure its quality through training and technical assistance.

A careful interpretation of the findings can lead to the following conclusions:

- The Wisconsin Long Term Care Functional Screen (LTCFS) is a reliable tool that can yield consistent results when administered to clients in similar circumstances by workers who receive appropriate training.
- The LTCFS is a valid tool for the establishment of Nursing Home Level of Care. Nursing Home and DD level of care obtained by applying the LTCFS logic will achieve high levels of agreement with level of care determination attained by utilizing Wisconsin's Bureau of Quality Assurance process and decision logic, when administered by properly trained workers.
- The documented efficacy of the LTCFS not withstanding, the instrument is a dynamic and evolving tool that can benefit from on-going quality assurance and improvement efforts. Such efforts will include: A. On-going training and technical support for the workers who will administer the LTCFS. B. Conducting small-scale studies and other quality control measures to assure the efficacy of the workers who engage in the screening process. C. incorporating new and tested elements, taken from the body of both practice and research.

REFERENCES

Altman, D. (1991). Statistics in Medical Research. New York: Chapman and Hall.

Campbell, D. and Stanley, J. (1967) Experimental and Quasi-Experimental designs for research. Chicago: Rand McNally.

Cochran, W. (1977). Sampling techniques. New York: John Wiley and Sons.

Cohen, J. (1960). Coefficient of agreement for nominal scales. Educational and Psychological Measurement, 10, 37-46.

Howell, D. (1992). Statistical methods for psychology. Belmont, CA: Duxbury Press.

Schuerman, J., Rzepnicki, T., and Littell, J. (1994). Putting families first. New York: Aldine De Gruyter.

Simon-Rusinowitz, L., Mahoney, K., Desmond, S., Shoop, D., Squillace, and Fay, R. (1997). Determining consumer preferences for a cash option: Arkansas survey results. Health Care Financing Review, 19 (2), 73-96.