Wisconsin Children's Long-Term Supports Functional Screen

Written as part of the Wisconsin Department of Health and Family Services' application to the Centers for Medicare & Medicaid Services for approval of the Children's Functional Screen for use by Wisconsin Children's Waivers

Authors: Ann Pooler, Nachman Sharon, Beth Wroblewski

Wisconsin's Functional Screen Method

The Wisconsin Children's Long-Term Supports Functional Screen (CLTS FS) has been developed by the Department of Health and Family Services workgroup in 2002 – 2003, as part of Wisconsin's Children's Long-Term Support Redesign project. The goal of the redesign project has been to improve access, coordination, choice, quality, and financing of the long-term support system to better serve children and families. The CLTS FS is to be completed by certified screeners through an interview with the child and parent(s), usually in the home. It has been programmed into a "one-stop shopping" electronic application that yields virtually instant eligibility results for multiple Medicaid and state- and county-funded programs for children with long-term supports needs.

The CLTS FS is built upon the success of Wisconsin's Long-Term Care Functional Screen (LTC FS) for adults, which since 1999 has been providing on-line eligibility determinations and specific levels of care (explained below) for Wisconsin adults with long-term needs (frail elders and people with physical and developmental disabilities). The LTC FS has proven accurate and reliable.[1]

The CLTS FS functions independently of the adult screen and is more complex because:

- It must work well for children from birth through young adult (age 22 years) with a variety of functional abilities, health-related needs, and disability types (physical, intellectual, or emotional disabilities, and/or mental illness);
- It determines eligibility for multiple programs having different eligibility criteria;
- It determines specific nursing home levels of care; and
- Specific diagnoses are required for eligibility for some programs.

The Department has numerous mechanisms to ensure that no applicant will be erroneously denied program eligibility through the CLTS FS. All denials will require review by a State nurse consultant.[2] Families will be notified of their appeal rights in accord with existing

[1] For more information, see https://www.dhs.wisconsin.gov/functionalscreen/index.htm.

[2] The adult functional screen does eligibility approvals and denials. Because the LTC FS reflects consensus of numerous experts, it is the best possible decision-making. In-depth analyses of complex cases since 1998, with some "fine-tuning" of the instrument, have proven the adult LTC FS to be a more reliable and accurate "gold standard" decision entity than individual state staff.

federal and state laws and policies. Screeners must meet educational and experiential qualifications and must complete an on-line screener training course and pass certification exams. Quality assurance and improvement exist at multiple levels from programmed entry-level edits and error cues, through mandatory local QA/QI processes, direct access to designated State consultants (nurses and social workers expert in the CLTS FS), to targeted, random, and individualized reviews and data analyses by State staff.

The components of the CLTS FS are as follows:

- **Demographics**, including information about county of residence and responsibility, living situation and medical insurance
- **Diagnoses**
- **Mental Health and Psychiatric Symptoms**
- **Behavioral Needs**
- **Activities of Daily Living (ADLs)** including age appropriate skills in bathing, dressing, grooming, mobility, transfers, eating, and toileting
- **Instrumental Activities of Daily Living (IADLs)** including**,** as appropriate for the child's age, communication, learning, meal preparation, and money management
- **Work and School** including information about the current school/work situation as well as supports needed and interests for future employment
- **Health-Related Services** including skilled nursing tasks, therapies, and other medically-oriented interventions

The CLTS FS is a needs inventory and not a complete strengths-based assessment. The following are the "screen development criteria" used to guide CLTS FS development:

- **Objectivity and Reliability**: The CLTS FS is designed to be as objective as possible in order to reach the highest possible **"inter-rater reliability"** (two screeners would answer the same way for a given child). Subjectivity must be minimized to ensure fair and proper eligibility determinations, as well as to improve statewide consistency.
- **Accuracy**: Eligibility determinations must be correct and must match current federal and state criteria in **every** instance.
- **Brevity**: The CLTS FS is only a "functional assessment" to determine program eligibility. It serves as a baseline for more in-depth assessment to develop a service plan that reflects each child's and family's strengths, values, and preferences.
- **Inclusiveness**: Children of all ages; with emotional, cognitive disabilities, physical disabilities, or developmental disabilities; with or without skilled nursing needs; in any setting from homeless to hospitals or institutions; can be accurately screened with the given choices for each question.
- **Clarity**: Definitions and answer choices, including diagnoses and nursing needs, must be clear to screeners with a broad array of professional backgrounds and experiences.

The CLTS FS determines functional eligibility for applicants from birth to age 21 years inclusive, and where relevant, a specific "Level of Care," (explained below) for five different programs:

- Katie Beckett Medicaid Program
- Family Support Program

- Community Options Program
- Mental Health Wraparound Services
- Children's Home and Community-Based Services Waivers for children with developmental disabilities or physical disabilities

The screener will collect relevant functional eligibility information in the course of meeting a child and their family. Again the CLTS FS is not a comprehensive assessment; rather it is a review of key information related to functional eligibility. Once the CLTS FS screen fields are complete, the computer eligibility logic is able to determine Hospital (HOS), Psychiatric Hospital (SED), Nursing Home (NH), and Developmental Disability (DD) Level of Care (LOC) for Home and Community-Based Services Waivers, as well as the Katie Beckett Medicaid Program. Additionally, the Target Group(s)--Physical Disability, Mental Health or Developmental Disability--for the Home and Community-Based Services Waivers is determined. The CLTS FS will also automatically indicate eligibility for other state- and county-funded long-term supports programs such as the Family Support Program, Community Options Program, and Mental Health Wraparound.

In addition to providing rapid program eligibilities, the CLTS FS is designed to:

1. Serve as a foundation for the comprehensive assessment related to long-term supports and services selected by the parent(s).
2. Provide data for quality assurance and improvement studies for the Department of Health and Family Services and long-term support programs utilizing the CLTS FS.
3. Provide data to counties and, as appropriate, to provider agencies on eligible children and on encounter data and timeliness of the eligibility process.

CLTS FS Development and Validity Testing

Development of the CLTS FS began in 2001 with the formation of a "screen development workgroup" of State eligibility decision-makers and others expert in children's long-term support programs. Over time these experts have added CLTS FS expertise to their repertoires, such that they can now "fine tune" the CLTS FS as needed so that it matches their expert consensus in every instance. The core workgroup continues to oversee CLTS FS development and testing, implementation, and quality assurance. Additional experts participated for months at a time to help develop assist with particular areas such as child development stages, mental illness, substance use problems, and employment.

Development of the CLTS FS was an iterative process of drafting, testing (on individual cases) and revising screen items and logics many times. While standard scale development processes were followed,[3] the requirement of accurate eligibility determinations for every

[3] See, e.g., Ian McDowell and Claire Newell, Measuring Health: A Guide to Rating Scales and Questionnaires, 2nd ed., NY: Oxford UP, 1996; J. Allen and Wendy M. Yen, Introduction to Measurement Theory. 2002. IL:Waveland Press; Robert F. DeVellis, Scale Development: Theory and Applications, 2nd ed, 2003, CA: Sage Publications; Richard G Netemeyer, William O. Bearden, Subhash Sharma, Scaling Procedures: Issues and Applications, 2003. CA: Sage Publications.

applicant required in-depth clinical analyses of individual cases in addition to the usual statistical analyses.[4]

Content validity—that the CLTS FS measures what it should, i.e., that it includes sufficient items appropriate to determining program eligibilities—was obtained initially through the development workgroup itself, since it was comprised of decision experts. Review for content validity was also solicited from stakeholders through numerous presentations of the CLTS FS to various county, provider, and advocacy groups. Content validity feedback was then elicited from screeners testing the CLTS FS, who were trained as "co-developers" and encouraged to provide feedback on all aspects of the CLTS FS and its instructions. (Such feedback is a critical part of on-going quality assurance for the screen.) While most CLTS FS content was derived from existing guidelines, other developmental criteria were procured from other sources such as child development guidelines.

Criterion validity—that CLTS FS items and eligibility results match "gold standards" of current federal and state eligibility criteria and expert decision making—was obtained through numerous methods. First, as noted above, screen development was an iterative process of drafting, testing, and revising items and logics based on workgroup members' analyses and applications to individual cases. This in essence comprises a first level of validity checking, comparing CLTS FS results with experts' decisions. The second level of criterion validity checking involved **concurrent reviews** in which State eligibility reviewers did side-by-side comparisons of the CLTS FS with their own decisions using current (mostly narrative) application forms. Initially, the State reviewers would manually complete a CLTS FS and check its eligibility results; later on, screen testers throughout the state submitted completed CLTS FS with application forms. Cases were discussed in regular meetings of State reviewers, notes from the concurrent reviews were saved in a database, and the screen development workgroup made revisions as needed. This concurrent review process began in 2003 and continues now. In effect, concurrent review will always be a part of on-going quality assurance and improvement, since every denial will be reviewed and since the screen development/oversight workgroup is comprised of decision experts. Unlike most research-based scales, Wisconsin's functional eligibility screens are dynamic and will be fine-tuned as needed to yield correct results for every single applicant, no matter how unique the situation.

Other forms of validity (predictive and construct validity, sensitivity, specificity, discriminant evidence) were developed for particular screen items and are discussed later.

[4] In fact, Wisconsin's functional eligibility screens are not merely scales, but automated expert decision instruments. Technically, the functional screens are not "expert systems" or "artificial intelligence" because they are not constructed with rules-based inference engine and a database. But functional screen scales and logics are complex and strategically overlapping enough to allow for uncertainty, fuzzy logic (categorical gradations, i.e., "grey areas") and lack of information. More importantly, the functional screen logics always yield the best possible decision—one reflecting consensus of expert decision-makers. For these reasons, the functional screens are in effect a form of artificial intelligence and decision-theoretic expert systems. See, e.g., Stuart J. Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, 2nd ed., 2003. NJ: Pearson Education, Inc.

CLTS FS Reliability

One of main requirements for the implementation and ongoing usage of a new measuring instrument is the establishment of its reliability commonly defined as the consistency and ability of the instrument to produce similar results under similar circumstances. In an effort to document the reliability of the Wisconsin Children's Functional Screen (hence CLTS FS), the Wisconsin Division of Disability and Elder Services in the Department of Health and Family Services had undertaken an extensive study between June and December of 2004. The findings of this study are described in this section.

To increase confidence in the findings of the study and in order to provide staff who are engaged in the continuous quality improvement of the CLTS FS with action oriented feedback, the study had simultaneously employed two complementing research designs: the Inter-Rater Reliability Study (IRRT), and a correct proportions test (explained below).

The IRRT is one of the leading designs for measuring reliability. The primary condition for its implementation is the ability to conduct repeated (cross) measurements of the same phenomena by two independent raters. The actual process involves the administration of an instrument (or key components thereof) to the same subject by two different raters within the same time frame. The stronger the agreement between the raters, the higher the confidence that the instrument is reliable (consistent). The design of our IRRT had an additional advantage by adding a component that enabled us to engage the two screeners in a follow up process of explaining the reasons for disagreements, as part of improving the quality of the CLTS FS (explained below).

The administration of the (same) instrument within the same time frame is an attempt to reduce the potential impact of maturation effects (Stanley and Campbell, 1967), e.g., to minimize differences that are due to objective changes in client condition between administrations. In the present study the instrument was administered and re-administered within two weeks and only if the child's condition had not changed significantly.

The IRRT assumes that both raters are adequately trained and that any differences in the ratings are not attributable to differences in screeners' competence and knowledge. Training was provided to all participating screeners.

Because repeated interviews of a child and family within a short time frame can be excessively intrusive, the design utilized a combination of limited live screenings of actual applicants, and cross-ratings by two screeners of written scenarios. This combined approach served to minimize the burden on families while retaining the essential characteristics of an IRRT design.

The second complementary design, correct proportions, was employed to validate the findings of the IRRT. It also provided staff, who is engaged in the continuous improvement of the screen and the screening process, with additional information about the performance of specific items in the screen under differing circumstances. The correct proportion test can be considered a variant of the IRRT. While the more conventional IRRT seek to measure

agreements between two raters, the correct proportion design captures the proportion of correct answers for each FS item within each written scenario. Correct answers were determined by consensus of the screen development workgroup who drafted and tested the written scenarios.

The Sample

The sample for the standard IRRT study included 190 randomly selected/assigned cases. Of these, 63 cases consisted of new applicants to the children's programs. The other 127 cases consisted of written scenarios that were read and rated (i.e., a simulated CLTS FS was completed on-line) by pairs of randomly assigned screeners. The final size of the sample (190) was chosen to assure the robustness of the statistical procedures and its makeup was meant to ensure sufficient representation of all relevant age and disability groups. Table 1 shows the distribution of sample cases among 11 age groups and disability cells:

Table 1
Distribution of Sample Cases Among Age and Disability Groups

Age Group (years)	DD	NH/Hosp LOC	SED/ MH	IEL	ILS	IAG	Total in Age Group
0-2	17	14	NA	2			33
3-7	17	11	19				47
8-13	8	22	15	1	2		48
14-21	21	20	11			10	62
	63	67	45	3	2	10	190

The fifteen cases that appear under the Ineligible columns (IEL, ILS, and IAG) have not been assigned to a disability type because they did not meet eligibility for any of the 3 waiver groupings—namely, developmentally disabled (DD), nursing home or hospital level of care (NH/Hosp LOC), or severe emotional disturbance or mental health problems (SED/MH). They were initially selected based on their age group and general assumptions regarding their best program fit. They were left, of course, in the final sample in spite of their "negative" screening outcome (lack of eligibility). There are no cases in the 0-2 SED/MH category because there are currently no acceptable screening and diagnostic tools that can reliably place children in this category.

The sample for the correct proportions' design consisted of 127 records and is a sub-sample of the larger IRRT sample. Each record included the scenario writer's ratings for each of the items included in the conventional IRRT, as well as the rating of an independent reader-rater. In all, there were 13 different scenarios that covered the different age, disability groups, and programs that are part of the Wisconsin children's Waivers.

A total of 57 unique screeners participated in the study. The number includes active professional staff from 16 Wisconsin counties, as well as State program staff and consultants.

CLTS FS Items Tested in IRRTs

The study analyzed levels of inter-rater agreement and correct proportions for sixty-four items that play a key role in program eligibility and assignment. Two secondary criteria were used to select the final items that were included in the analysis. The first was the relative importance of each item in the (automated) decision logic that determines program eligibility and disability type. The second criterion meant to ensure sufficient representation of the various domains that comprise the screen. The final list included items that measure functionality (activities of daily living [ADLs] and instrumental activities of daily living [IADLs]), items that measure the need for medical services and organ transplants (Health Related Services [HRS]), and items that gauge the risk for criminal justice services, school expulsion, and protective services. Additional items included several that identify children with violent, offensive and suicidal behaviors, eating disorders, and items that focus on behavioral, learning, and communication deficits.

Statistical Procedures

Three statistical measures were used by the researchers to establish the significance of agreement between the two screeners in the IRRT. The Cohen's Kappa (Cohen, 1997; Howell, 1992; Schuerman, Rzepnicki, & Littell, 1994) is commonly used to calculate correlation between the scores of two screeners. The Kappa coefficient has been amply documented and used extensively to measure agreement between screeners in a variety of fields. Examples include the measurement of agreement between husbands and wives' perceptions of shared events, between employers and employees on a desired outcomes, and between physicians (radiologists) 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 applications for both mainframe and microcomputers.

The Kappa coefficient is highly sensitive to seemingly minute levels of disagreements between screeners. Thus, even when simple cross-tabulations point to high level of association/agreement, Kappa coefficients can in some instances look artificially low not because of the number of disagreements, but rather because of the way the disagreements are distributed among the cells. For those items where cross tabulations and writer-screeners correct proportions reflect high agreement levels, and where the value of Kappa was 0.72 or higher, the measure was the coefficient of choice. Where cross tabulations and writer-screeners correct proportions reflect high agreement levels, but the value of Kappa was less than 0.70 indicating only fair to good agreement beyond chance (Fleiss, 1981), the researchers opted to calculate an alternative coefficient, the Gamma.

It should be noted that the researchers' decision to employ additional measures in certain cases reflects a cautious and conservative approach to the analysis. Although the Kappa benchmarks for acceptance depend primarily on the level of risk associated with making erroneous decisions, Landis and Koch (1977) have suggested the following coefficient benchmarks as general guidelines: 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 noted earlier, this research has established a minimum Kappa threshold of 0.72.

The Goodman-Kruskal Gamma is one of the most commonly used non-parametric tests for measuring correlation between independent scores for ordered (ordinal) variables such as most items on the CLTS FS (Hays, 1981). In general, the existence of an attribute (or higher value) on a screen item contributes to higher program eligibility and vice versa. The Gamma is particularly useful when many of the scores are tied (for example when most or all scores are either one or zero), and the Statistic is considered preferable to the Spearman's Rho and to the Kandall's Tau under such conditions. The Goodman-Kruskal Gamma is also documented and included in both the SAS and SPSS applications.

The third test used for all the items as part of the IRRT (as applicable) was the Chi-square test of association, used to establish whether the distribution of agreements and disagreements between the screeners is attributable to chance.

For all three measures, coefficients were judged to be statistically significant when the p. value (the probability that a specific configuration of responses that was detected by the study could have occurred by chance) was less the 0.05 (or 5%).

Technical Aspects of Data Gathering and Analysis

Children's Functional Screen IRRT testing was performed with the assistance of a variety of supporting technologies. These technologies allowed us not only to accelerate the testing timeline, but also facilitated the subsequent data gathering, aggregation, and analysis.

The researchers took advantage of the FSIA application to collect the test data. Supported on Microsoft Windows 2003 servers, FSIA is a web-based application written in ASP/COM coding language and protected by a firewall to block any unauthorized access. The use of JavaScript as the application's scripting language aided in the data integrity process by creating error and warning messages at the time of data entry.

FSIA has two regions that interface with the screeners - the Production and the Training regions. Since the application in the Training region mirrored Production, we were able to use both regions to collect test data. When actual applicant screen information was used, the first rater-screener entered the data into Production and the second rater-screener entered the data into the Training region. When test case scenarios were used, both primary and the secondary screeners entered the data into the Training region.

To aggregate and analyze results, the data was first extracted via XML file protocol from an Oracle database. The extracted data was then imported into Microsoft Access where the results from the primary and secondary screeners were paired-up. The resultant data set was then imported into an SPSS application for statistical analysis.

Findings

The following Table 2 provides information on the four procedures performed on the 64 items that were included in the study: the proportions of writer-rater agreements for items in written scenarios, the Kappa or Gamma correlation coefficients for inter-rater level of agreement, and the Chi-square measure of association. For the last three measures, levels of statistical significance are also presented.

In Table 2, the first column in the table gives the proportions of ratings that are correct (percentage of writer-rater agreements) for each item in the written scenarios. The second column provides an item description. The third column shows the IRRT correlation coefficient (Kappa or Gamma as appropriate) for the item. The forth column provides the level of statistical significance of the IRRT correlation coefficient. The fifth column displays the raw Pearson Chi-square score for the item. The sixth provides the number of degrees of freedom (df) based on the number of cells in each item specific cross-tabulation. The last column shows the level of statistical significance of the Chi-square measure, based on the raw Pearson score and the number of degrees of freedom.

Table 2

Correct Proportions and Inter-Rater Correlations for CLTS FS

Correct Proportion *N= 127*	**CLTS FS Item**	**Inter-Rater Reliability Test** *(All cases, N=190)*				
		Kappa/ Gamma	**p <**	**Chi Square**	**df**	**p <**
ADLs and IADLs						
97.6%	Bathing Need	0.835	0.00	132.4	1	0.00
93.7%	Dressing-Need for Help	0.725	0.00	100.2	1	0.00
95.3%	Eating-Need for Help	0.723	0.00	99.4	1	0.00
98.4%	Mobility-Need for Help	0.946	0.00	169.9	1	0.00**
94.5%	Toileting-Need for Help	0.894	0.00	152.6	1	0.00**
97.6%	Transferring-Need for Help	0.922	0.00	161.7	1	0.00**
96.9%	Grooming-Need for Help	0.797	0.00	120.8	1	0.00
99.2%	Meal Preparation Need-None Applied	0.945	0.00	170.0	1	0.00**
100.0%	Needs Assistance to Work	UTC*	NA	UTC*	NA	NA
100.0%	Money Management Need-None Applied	1.000	0.00	190.0	1	0.00**
Behavioral Issues (Check all that apply)						
85.0%	Lack of Behavioral Controls	0.773	0.00	40.3	4	0.00**
91.3%	Failing Grades/ Truancy	0.958	0.00	74.3	1	0.00
92.9%	High Risk Behaviors	0.763	0.00	229.2	4	0.00**
92.1%	No Behavior Problems	0.848	0.00	137.4	1	0.00
76.4%	Social Roles and Interactions	0.864	0.00	55.3	1	0.00
93.7%	Self-Injurious Behaviors	0.925	0.00	55.0	1	0.00

Correct Proportion *N= 127*	CLTS FS Item	Inter-Rater Reliability Test *(All cases, N=190)*				
		Kappa/ Gamma	p <	Chi Square	df	p <
96.9%	Aggressive or Offensive Behaviors	0.939	0.00	77.5	1	0.00
90.6%	Violence	0.976	0.00	96.6	1	0.00
89.8%	Child Needs n-School Supports for emotional or behavioral problems	0.889	0.00	58.7	1	0.00
Health-Related Services (check all that apply)						
96.9%	Misses Over 50% of School due to health	0.759	0.00	109.9	1	0.00
100.0%	Emergency Medical Incidents With a Sudden Onset	UTC*	NA	UTC*	NA	NA
97.6%	Bowel-related Skilled Tasks-Digital Stim, Ostomy site care	0.953	0.00	69.8	6	0.00**
100.0%	Dialysis	UTC*	NA	UTC*	NA	NA
97.6%	IV Fluids or Medications	0.990	0.00	385.1	16	0.00**
100.0%	Suctioning	1.000	0.00	190.0	2	0.00**
100.0%	Respiratory Treatments	1.000	NA	190.0	2	0.00**
96.1%	Does not Need Tube Feeding	0.972	0.00	149.9	4	0.00**
96.1%	Urinary Catheter-related skilled tasks	0.957	0.00	90.7	4	0.00**
100.0%	Total Parenteral Nutrition (TPN)	0.898	0.00	30.5	6	0.00**
99.2%	Transplanted Organ-Bone Marrow	0.939	0.00	167.4	1	0.00**
100.0%	Transplanted Organ-Heart	UTC*	NA	UTC*	NA	NA
100.0%	Transplanted Organ-Intestine	UTC*	NA	UTC*	NA	NA
99.2%	Transplanted Organ-Kidney	0.954	0.00	173.1	1	0.00**
100.0%	Transplanted Organ-Liver	UTC*	NA	UTC*	NA	NA
100.0%	Transplanted Organ-Lung	UTC*	NA	UTC*	NA	NA
100.0%	Transplanted Organ-Pancreas	UTC*	NA	UTC*	NA	NA
92.9%	Impairment at Least One Year Prior	0.923	0.00	58.5	1	0.00
96.1%	Positioning Every 2 Hours	0.978	0.00	71.4	1	0.00**
98.4%	Recurrent Cancer Not in Remission	0.954	0.00	173.2	1	0.00**
100.0%	Brain Injury Rehabilitation Program Minimum 15 Hrs/Wk	UTC*	NA	UTC*	NA	NA
78.8%	Nursing Needs Expected to Last	0.977	0.00	127.3	4	0.00**
70.9%	Nursing Needs Have Lasted	0.908	0.00	169.7	4	0.00**
100.0%	Terminal Condition	UTC	NA	UTC*	NA	NA
94.5%	PT, OT, or ST by Therapist	0.749	0.00	107.5	1	0.00
95.3%	PT, OT, or ST follow-through by others	0.936	0.00	88.7	1	0.00
100.0%	Ventilator	1.000	0.00	190.0	1	0.00**
96.1%	Wound or Special Skin Care	0.871	0.00	16.4	1	0.00**
97.6%	Learning Impairments-None Applied	0.842	0.00	134.7	1	0.00
98.4%	Has Persistent Emotional Disability	0.855	0.00	139.3	1	0.00
Child has the following symptoms of mental illness						
59.8%	No Symptoms of Mental Illness	0.293	0.05	3.8	1	NS

Correct Proportion *N= 127*	CLTS FS Item	Inter-Rater Reliability Test *(All cases, N=190)*				
		Kappa/ Gamma	**p <**	**Chi Square**	**df**	**p <**
100.0%	Psychosis	0.954	0.00	173.2	1	0.00**
99.2%	Suicidality	0.849	0.00	137.2	1	0.00**
100.0%	Anorexia	UTC	NA	UTC*	NA	NA
Child currently requires the following services						
100.0%	Criminal Justice System	0.938	0.00	168.0	1	0.00**
100.0%	Child Protective Services	UTC*	NA	UTC*	NA	NA
95.3%	Mental Health Services	0.837	0.00	133.6	1	0.00
93.7%	Substance Abuse Services	0.956	0.00	47.9	1	0.00
93.7%	In School Supports for Emotional/Behavioral Problems	0.765	0.00	111.3	1	0.00
96.9%	Services Required more than 3 Times/Week	0.749	0.00	106.8	1	0.00
Child has no impairments now but has confirmed diagnosis expected to lead to impairments within 1 year in any of the following						
79.5%	Functional Impairment Expected Within 1 Year--Communication	0.578	0.00	14.5	1	0.00
83.5%	Functional Impairment Expected Within 1 Year-Learning	0.604	0.00	15.1	1	0.00
78.8%	Functional Impairment Expected Within 1 Year-Mobility	0.789	0.00	33.5	1	0.00
69.3%	Functional Impairment Expected Within 1 Year-Self-Care	0.683	0.00	25.8	1	0.00

* UTC Unable To Calculate
** Chi Square Significance is Somewhat Overstated

Of the sixty-four items that were analyzed, forty-eight had met or exceeded our benchmark of 0.72: ten achieved a coefficient of 0.72 to 0.80, twelve had scores of between 0.80 and 0.90, and twenty-six items were in the 0.90 to 1.00 range. Four items failed to meet our minimum acceptance trash-hold. Three in the area of functional impairments and one in the area of mental health symptoms. The accepted coefficients are all statistically significant beyond the p. <0.05 level. In several of the Chi-square tests, statistical significant is somewhat overstated because several of the cells contain fewer than 5 cases. A cautionary note is added next to each of these measures.

The statistical program was unable to calculate (UTC) valid coefficients for twelve of the items, due to lack of variation among the responses. For these items, the scores of both first and second rater were all concentrated in a single answer cell. For example, in the area of organ transplant, all the responses without exception fell into the zero cells (e.g., all the screeners agreed that the applicant did not have that medical condition). One can say that the

program inability to calculate a coefficient for these items is purely technical and in effect agreement was very high between the screeners. This argument can be supported by the findings of high levels of writer-rater agreements (correct proportions) in written scenarios for most of these items.

The supplementary methodology of calculating the number of agreements between the writers of the simulated scenarios and independent readers seem to validate and confirm the findings of the conventional IRRT. It also seems to reinforce earlier explanations that measures of correlation have certain limitations, which result from the need to meet certain conditions related to cell distributions. Nevertheless, as can be seen from the table, there is a general pattern of agreement among all the measures used in the study.

Clinical Analyses and Improvements

In addition to the statistical analyses explained above, more detailed "clinical" feedback on the reasons for certain discrepancies was solicited from screeners. For each pair of IRRTs, a "Discrepancy Reasons Report" was generated that listed the differences between the two screens. Screeners were asked to meet to review this to discuss and record the reasons for each discrepancy. To simplify, a list of coded options was presented as follows:

1. A mistake by a screener (knew better, but erred)
2. A misunderstanding by a screener
3. Unclear wording on the screen, or choices don't "fit" well
4. Unclear training on the screen
5. Different perceptions or interpretations by screeners
6. Inconsistent info provided, unspecified by whom
7. Different info given by consumer at different times
8. Different info given by consumer and others (staff, family, caregivers)
9. Condition changed

Conference calls were held with IRRT screeners of different agencies to discuss IRRT discrepancies and feedback. All screeners were encouraged to contact State staff with feedback. These processes provided invaluable insights to guide revisions to the CLTS FS, logic, and instructions.

In fact, all "weak" areas in the CLTS FS—those with lower correlation scores shown above—were identified long before IRRT data results were available, through case analyses and user feedback elicited early in the IRRT process. Proposed revisions were discussed with screeners, thus adding their expertise directly to the screen development process. Consensus was reached that each revision would improve a problematic item or section on the CLTS FS. IRRT statistical results confirmed developers' and users' analyses of items needing revision. Some key revisions and improvements are described in the next section.

Analyses of Items with Lower IRRT Scores

While Kappa or Gamma scores of 0.70 to 0.90 are statistically acceptable, they are not clinically as precise as desired.

ADLs and IADLs

The IRRT process exposed several misunderstandings that led to lower IRRT scores for some ADLs. Several screeners scored a child based on their one-visit observation alone, rather than using all available information sources. Other screeners learned through the IRRT that more precision is needed than they had employed. For instance, one screener marked a child independent with toileting because mom answered "Yes" to the question "Is she potty-trained?" when in fact the other screener determined that the child still needs help due to her disability.

Bathing, dressing, and grooming were revised to reduce subjectivity regarding the need for help and to clarify the need for adaptive equipment. Some answer choices for specific age groups were revised to reduce "grey areas" (subjectivity).

Health-Related Services

Reliability scores were good for each of the health-related conditions or tasks and for the **duration** of the skilled nursing needs. (Lower correct proportions on the duration questions are an artifact of the +correct answer distinguishing 6 months and 12 months or more.)

Mental Health and Behavioral Challenges

The very low reliability scores for "**child has no symptoms of mental illness**" resulted mostly from an artifact of the FS computer application itself: When screeners left the mental health/behaviors page, the application auto-filled the mental health symptoms to "none." This error was corrected. Beyond this, however, IRRT trainings and screener feedback did anticipate the low reliability scores for some of the mental health items. In particular, screeners were used to relying on diagnoses alone and were unclear on whether particular behavior problems constituted "symptoms of mental illness." Some screeners were not comfortable asking parents about mental illness symptoms. Additional instructions were written to clarify these items and their importance, and this area will be closely monitored and additional training developed to ensure correct eligibility results.

The CLTS FS layout caused some confusion between "mental illness" and more general behavioral problems, both of which were on the same page. Current eligibility criteria for mental health programs include violence as a symptom of mental illness. More broadly, however, violence is also important in considering eligibility and needs for children with cognitive, developmental, or emotional disabilities but without mental illness. "Violence" appeared in both sections of the CLTS FS, which created confusion for screeners, as reflected in low IRRT scores for these items. This problem was corrected in the screen and instructions.

Several behavioral items, particularly "**lack of behavioral controls**," had lower reliability mostly because there was some confusion on whether screeners should check only one or several behavioral challenges listed on the FS. For instance, some screeners checked "lack of behavioral controls" because they had checked "suicidal" for the child. Trainings and instructions were revised to clarify this. Revisions were also made to reduce subjectivity.

"**Social roles and interactions**" is a particularly difficult item to capture objectively and is expected to have lower reliability than any other screen items. It is included because it can be an important eligibility consideration for children who otherwise function well and lack behavioral problems. The question was completely rewritten during the IRRT process and discussed with many screeners who concurred that the newer version would improve objectivity and reliability. Its validity and reliability will be monitored as part of FS quality oversight.

"**Child currently requires the following services…**" had low reliability for several reasons. 1) Screeners were unsure whether to check it if the child was not receiving any services but should be (e.g., if parents declined or couldn't afford services or if services were unavailable); 2) It was not clear how screeners were to ascertain such needs; and 3), The definitions of the services were not clear, so screeners were unsure what counted. All of these issues were addressed in discussions with screeners and the items and instructions were revised accordingly.

"**Child needs in-school supports for emotional/behavioral problems**" appears twice on the CLTS FS, with 89.8 and 93% correct proportion scores and Kappas of .889 and .765. Based on screeners' feedback, additional instructions were added concerning active vs. inactive behavioral interventions plans and what constitutes "supports" for this item.

"**Child has no impairments now but has confirmed diagnosis expected to lead to impairments within 1 year in…learning, communication, etc**." This group of items had low reliability because the leading question was confusing; it has been corrected. This item is important for the rather infrequent times when eligibility is appropriate because a high-functioning applicant's functioning will deteriorate significantly and rapidly over just a few months. (An example would be a fast-growing brain cancer known to cause motor impairments.)

Specificity and Sensitivity with Intellectual Disabilities

It was noted above that the CLTS FS accurately determines the disability type(s) for each applicant—physical disability, developmental disability, mental illness, and serious emotional disability. A more fascinating challenge has been for the CLTS FS to distinguish children meeting the federal definition of developmental disability from those with learning disabilities, dyslexia, attention deficit disorder, or mild delays possibly misdiagnosed as autism, autism spectrum disorder, and pervasive developmental disability, but with normal intelligence. Through a mix of diagnostic categorization and functional data including behavior, communication, learning, relationships, ADLs, and IADLs, the CLTS FS can now

distinguish children with intellectual disabilities matching the federal definition of "developmental disability" from those who do not—even if the same diagnoses are presented. Because it reflects expert consensus, the CLTS FS logic in matters such as this is more sophisticated and accurate than most practitioners, and is certainly more consistent across time and settings.

Conclusions

With the few item exceptions discussed above, Wisconsin's CLTS FS is a reliable eligibility tool that yields consistent results for long-term support program eligibilities, level of care, and type of disability, when administered in similar circumstances within the same time frame by workers who receive appropriate training. As also evidenced above, the Department is committed to on-going quality assurance and improvement, so that eligibility determinations are accurate for every applicant.

The documented efficacy of the Wisconsin Children's Functional Screen notwithstanding, the instrument is a dynamic and evolving tool that can benefit from on-going quality assurance and improvement efforts. Such efforts will include: 1) On-going training and technical support for screeners; 2) On-going testing and other quality control measures; 3) Fine-tuning of the FS instructions; and, 4) A web course based on lessons learned from the field and changes in practice and research. The parent agency of the Wisconsin children's waivers, the Division of Disability and Elder Services, already has an extensive infrastructure of databases and applications in place, to ensure the ongoing discovery, remediation, and improvement that are needed as part of the CLTS FS implementation.

REFERENCES

Allen, J. and Yen, W.M. (2002). Introduction to Measurement Theory. IL:Waveland Press

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.

DeVellis, RF (2003) Scale Development: Theory and Applications, 2nd ed. CA: Sage Publications

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions. New York: John Wiley.

Hays, W. (1981). Statistics. New York: CBS College Publishers.

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

McDowell, I and Newell, C. (1996) Measuring Health: A Guide to Rating Scales and Questionnaires. 2nd ed., NY: Oxford University Press

Netemeyer, RG, Bearden, WO, Subhash, S. (2003) Scaling Procedures: Issues and Applications, CA: Sage Publications

Russell, SJ and Norvig, P (2003). Artificial Intelligence: A Modern Approach. 2nd ed. NJ: Pearson Education, Inc.

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.