well SCREENING®

EXAMINER'S MANUAL

RESEARCH EDITION

Barbara L. Ekelman
Debra A. Dutka
Karen St. Amour

well
SCREENING®

CHAPTER 8

Technical Information

This chapter provides technical details on the development and norming of the Well Screening. The Well Screening is a valid and reliable tool developed through 30 years of clinical and research practice and 5 years of data collection and statistical analysis. All data were collected by professionals with at least master's degrees in speech-language pathology or education.

NORMATIVE SAMPLE

The data set for the Well Screening was collected by licensed speech-language pathologists and learning specialists over a period of 5 years from October 2014 through May 2019. A representative sample of kindergarten students that roughly reflects the demographic diversity of the suburban Midwest was sampled. Table 8.1 shows the kindergarten demographics by time of testing. Table 8.2 presents demographic information for Shaker Heights, Ohio; Cuyahoga County, Ohio; the entire state of Ohio; and the United States for comparison purposes. When comparing Tables 8.1 and 8.2, the Well Screening sample is roughly comparable to United States gender and race demographics. Unfortunately, socioeconomic status (SES) data was not available for the Well Screening sample. However, the geographic area from which the data was collected (Cuyahoga County, Ohio) fit within the SES parameters of the United States.

Table 8.1. Well Screening preschool and kindergarten demographics by time of screening

Assessment period	n	% Female	% Male	% White	% Black	% Hispanic	*% Other
K-Fall	274	56.57	43.43	69.34	10.95	2.55	17.15
K-Winter	269	56.13	43.87	69.14	11.15	2.60	17.10
K-Spring	269	56.13	43.87	69.14	11.15	2.60	17.10
Pre-K- Spring	189	61.38	38.62	73.02	7.41	3.70	13.23

*Other = Two or more races, Asian, Asian Indian, or Middle Eastern

Table 8.2. Demographic information for the total population in Shaker Heights, OH; Cuyahoga County, OH; Ohio, and the United States

Location	% Female	% White	% Black	% Hispanic	% Persons in poverty
Shaker Heights, OH	54.70	56.50	33.20	2.60	8.60
Cuyahoga County, OH	52.30	63.60	30.50	6.20	18.10
Ohio	51.00	81.90	13.00	3.90	14.00
United States	50.80	76.50	13.40	18.30	12.30

Table 8.3. Well Screening age by grade and time period

Age	Fall			Winter			Spring		
	min	max	average	min	max	average	min	max	average
K	4.99	6.42	5.71	5.27	6.70	5.99	5.56	7.03	6.30
Pre-K	n/a	n/a	n/a	n/a	n/a	n/a	4.43	6.00	5.22

n/a = not applicable; pre-K not tested in fall or winter

A total of 274 kindergarten students were screened with the Well Screening tool at the beginning of the school year for Assessment Period 1 (K-Fall). A total of 269 of the 274 children were again screened during the middle of kindergarten for Assessment Period 2 (K-Winter), and at the end of the school year for Assessment Period 3 (K-Spring). Additionally, the Well Screening was administered to 189 prekindergartners at the end of the school year (pre-K-Spring). Table 8.3 shows the age ranges and means for the subjects at each time period.

ITEM DEVELOPMENT AND TESTING

The Well Screening is a unique comprehensive screening drawn from the specialized fields of speech-language pathology, education, psychology, and child development, and is backed by more than 30 years of research and clinical experience. The creation of the items included in the Well Screening was based on developmental charts, interdisciplinary gold standard tests, and extensive review of the research on child language and learning development. Test items were crafted to tap into areas representing a broad range of skills important for school success: language, early literacy, reading, attention, math, social communication, speech sound production, and motor skills. See the section titled How the Well Screening was Developed in Chapter 1 to learn more.

EVALUATION ACROSS GROUPS

Linear Regression

Linear regressions were run on the data to evaluate the role gender played in test performance. Linear regression attempts to model the relationship between two variables (in this case, male and female gender) by fitting a linear equation to observed data. Correlation coefficients and p-values of the linear regression are shown in Table 8.4.

Table 8.4. Well Screening gender coefficients and p-values by subtest

Subtest		1	2	3	4	5	6	7	8	9	10
K-Fall	R^2	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00
	p-value	0.30	0.36	0.55	0.10	0.85	0.21	0.61	0.30	0.39	0.36
K-Winter	R^2	0.00	0.00	0.00	0.01	0.00	0.04	0.01	0.00	0.00	0.00
	p-value	0.87	0.28	0.75	0.07	0.95	0.02	0.15	0.55	0.73	0.95
K-Spring	R^2	0.01	0.01	0.00	0.00	0.00	0.07	0.00	0.00	0.01	0.01
	p-value	0.35	0.22	0.98	0.74	0.75	0.00	0.51	0.32	0.11	0.17
Pre-K-Spring	R^2	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.01	0.02
	p-value	0.69	0.67	0.06	0.24	0.52	0.77	0.75	0.29	0.27	0.06

Values in red are statistically significant; 1 = Language Processing; 2 = Number Sequences; 3 = Word Sound Play; 4 = Confrontational Naming; 5 = Pragmatics; 6 = Calculation; 7 = Language Formulation (morphology); 8 = Letter Recognition; 9 = Real Word Reading; 10 = Nonsense Word Reading.

As illustrated in Table 8.4, gender does not play a significant role in any of the subtests in prekindergarten. The p-value remains larger than 0.05 for all subtests, and the correlation coefficients remain low. When looking at kindergarten, there are no statistically significant differences during K-Fall. During K-Winter, there is a statistically significant gender difference (p=0.02) for Subtest 6, Calculation. This difference continues into K-Spring (p=0.00). The items in Subtest 6 measure math skills, and the data show that boys outperform girls in math at K-Winter and K-Spring. These gender trends are consistent with studies that have examined children's trajectories of mathematic achievement using participants from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) with boys outperforming girls in math as early as kindergarten (Cornwell et al., 2013; Wei et al., 2015).

Differential Item Analysis (DIF)

In order to be sure that each item in each subtest was fair and not introducing gender or racial bias, a Differential Item Analysis (DIF) was performed. It is important to ensure that items are of similar difficulty across groups of children. The Mantel-Haenszel statistic, which uses a Chi Square contingency table approach, was used to determine if bias existed. This analysis was performed on nine of the ten subtests that are scored as correct or incorrect. The analysis was not calculated for Subtest 4, Confrontational Naming, because it is a timed subtest that is scored in seconds rather than as a correct or incorrect response.

The data were analyzed by controlling for subtest difficulty. Table 8.5 shows the items and time periods where bias was identified based on gender. Sixteen items out of a total of 424 items (4 x 106) were found to show gender bias at one time period. Out of the sixteen items showing gender bias, eight items were biased toward boys and eight items were biased toward girls. Therefore, none of the items were eliminated based on the gender-balanced results of this analysis.

Table 8.5. Frequency of differential item functioning by subtest and time period by gender

Subtest	1	2	3	4	5	6	7	8	9	10
Number of items	12	10	10		10	14	14	12	12	12
K-Fall	0	0	0		0	Q8^{m}	Q4^{f}	Q8^{f}	Q6^{m}	0
K-Winter	0	Q3^{f}	0		0	Q6^{m}	0	0	0	0
K-Spring	0	0	0		0	Q7^{m} Q12^{m}	Q8^{f}	0	Q7^{f}	0
Pre-K-Spring	0	0	0		Q2^{m} Q3^{f}	Q8^{f} Q14^{m}	0	Q7^{m} Q8^{f}	0	0

Key: Q = Question; M = Male; F = Female
Note: Differential Item Analysis was not calculated for Subtest 4 because this subtest is scored in seconds and not as a correct/incorrect response.

Table 8.6 shows the items and time periods where bias was identified based on race. In this analysis, sixteen items out of a total of 424 items (4 x 106) were also found to show racial bias at one time period. Ten of the sixteen items were biased toward white children, while the remaining six were biased toward children of other races. None of these items overlapped across time periods. Therefore, it was concluded that no items needed to be eliminated based on the results of this analysis.

Table 8.6. Frequency of differential item functioning by subtest and time period by race

Subtest	1	2	3	4	5	6	7	8	9	10
Number of items	12	10	10		10	14	14	12	12	12
K-Fall	0	0	Q7^{nw}		0	Q3^{nw}	Q6^{w} Q8^{w}	0	0	Q3^{nw}
K-Winter	0	0	Q6^{nw} Q9^{w}		Q1^{w} Q3^{w}	0	0	Q11^{nw}	0	0
K-Spring	0	0	0		0	0	0	0	0	Q11^{nw}
Pre-K-Spring	0	Q3^{w}	0		0	0	Q8^{w}	Q1^{w} Q2^{w} Q3^{w}	0	0

Key: Q = Question; W = White; NW = Non-White
Note: Differential Item Analysis was not calculated for Subtest 4 because this subtest is scored in seconds and not dichotomous.

RELIABILITY

Reliability can be defined as the extent to which an individual's score (deviated from the mean) remains consistent over repeated administrations of the same test. If a student takes the same test repeatedly the results should be consistent. There are many approaches for estimating reliability scores. The test–retest method was chosen because the same students were being administered the screener at multiple time points. It should be noted that some variability is expected because the students are learning information in school at varying rates. The test–retest method of measuring reliability is the most effective for measuring constructs like intelligence.

Internal Reliability

For data to be reliable, it is important to make sure that each variable within each test is consistent for what it is measuring. To assess the internal reliability of the data, the Kuder-Richardson 20 (KR20) formula was used. This is a measure of internal consistency that is used when assessing data that is binary or dichotomous in nature. Because all of the subtests except for Subtest 4, Confrontational Naming, are of a dichotomous nature (the student's response is either correct or incorrect), KR20 is the appropriate statistic to use. Subtest 4 is a timed task measuring the length of time it takes the student to name a series of common objects. The KR20 coefficient ranges from 0 to 1.0, where values less than 0.39 are considered poor; 0.40 to 0.59 are adequate; 0.60 to 0.79 are good; and 0.80 to 1.0 are excellent.

Table 8.7. Internal reliability Kuder-Richardson 20 values by grade and subtest

Subtest	1	2	3	4	5	6	7	8	9	10
K	0.44	0.66	0.80		0.41	0.76	0.47	0.77	0.93	0.91
Pre-K	0.37	0.68	0.77		0.39	0.72	0.61	0.85	0.92	0.85

Key: 1 = Language Processing; 2 = Number Sequences; 3 = Word Sound Play; 5 = Pragmatics; 6 = Calculation; 7 = Language Formulation (morphology); 8 = Letter Recognition; 9 = Real Word Reading; 10 = Nonsense Word Reading
Note: Internal reliability was not calculated for Subtest 4 because this subtest is scored in seconds and not dichotomous.

As shown on Table 8.7, the kindergarten values ranged from 0.41 (Subtest 5) to 0.93 (Subtest 9). The prekindergarten values ranged from 0.37 (Subtest 1) to 0.92 (Subtest 9). It appears from these data that Subtest 5, Pragmatics, is not very consistent; however, upon careful examination, none of the questions were significantly (or even slightly) problematic. When performing a KR20 analysis, if there are problematic questions, it is easy to identify the problem items by removing them and thus increasing the KR20 value. However, this was not the case for Subtest 5, and it was concluded that the lower KR20 value is most likely a function of the variability of young children's social development and the linguistic and nonlinguistic complexity of the subtest. Subtest 5 requires the integration of language processing, knowledge of nonverbal cues, and problem-solving skills based on contextual cues. Children's skills vary among these factors, which may also explain the lower KR20 value. Further, when comparing K-Fall to K-Spring, performance dramatically improves from 37% of the children getting all of the questions correct to 65% of the children getting all of the questions correct. Furthermore, a simplified analysis of the predictive power revealed that only 1% of children identified as at-risk during K-Fall were at risk at K-Winter. This furthers the belief that kindergarten is just a period of time during which complex social communications skills are developing.

The results across the items in Subtest 1, Language Processing, were also not consistent, even though similar items on other standardized tests have shown consistency (*Clinical Evaluation of Language Fundamentals–Preschool–2,* Semel et al., 2004; *Preschool Language Scale,* 5th ed., Zimmerman et al., 2011). The inconsistency found in Subtest 1 may be due to differences in the time needed for each child to acclimate to the screening process or variability in active listening skills, which, in turn, interfere with the child's ability to focus on the task at hand.

Test-Retest Reliability

Test–retest reliability measures whether or not the same results will occur when a test is administered again to the same student. Pearson's *r* was used as a measure of test–retest reliability of the students when tested at K-Fall and K-Winter. Given the amount of time between testing (approximately 3 months) and maturational growth occurring, the correlations are not expected to be especially high. Correlations between time periods were therefore evaluated using these standards: Trivial (0.0 to 0.19), Low (0.20 to 0.39), Moderate (0.4 to 0.59), Strong (0.60 to 0.79), Very Strong (0.80 to 0.99), and Perfect (1.00). Coefficients that were at least in the moderate range (0.40 or better) are considered to be acceptable.

As shown in Table 8.8, the correlation coefficients ranged from a low of 0.37 (Subtest 5) to a high of 0.69 (Subtest 9). The reliability of nine of the subtests is acceptable when taking growth into account and the large variability within the kindergarten population because of their young age. The correlation coefficient for Subtest 5 falls below the acceptable range; however, as explained previously, due to the variability in young children's social development, the subtest is still considered reliable.

Table 8.8. Test-retest reliability Pearson coefficient for Well Screening K-Fall-Winter

Subtest	1	2	3	4	5	6	7	8	9	10
Pearson's *r*	0.55	0.66	0.68	0.62	0.37	0.66	0.41	0.43	0.69	0.67

1 = Language Processing; 2 = Number Sequences; 3 = Word Sound Play; 4 = Confrontational Naming; 5 = Pragmatics; 6 = Calculation; 7 = Language Formulation (morphology); 8 = Letter Recognition; 9 = Real Word Reading; 10 = Nonsense Word Reading

VALIDITY

One of the most important components to consider when evaluating a psychometric test is the validity of the instrument. Validity refers to how accurately a method measures what it is intended to measure. The validity of the Well Screening was examined by checking how well the Well Screening results corresponded to established results of gold standard tests.

In order to assess the validity of the Well Screening, 2 x 2 contingency tables were constructed to compare the Well Screening cutoff values to nationally normed tests (see Table 8.9). In addition to the Well Screening, 87 of the kindergarten students were also administered the *Kindergarten Diagnostic Instrument (KDI-II)* (Miller, 2000). The KDI-II measures developmental readiness over thirteen different areas. Another 89 kindergarten children were also assessed with the *Kaufman Brief Intelligence Test, Second Edition (KBIT-2)* (Kaufman & Kaufman, 2014), which measures verbal and nonverbal cognitive ability. Additionally, 89 children were administered the *Kaufman Survey of Early Academic and Language Skills (K-SEALS)* (Kaufman & Kaufman, 1993), which measures children's language skills (receptive and expressive), pre-academic skills, and articulation.

The results of the Well Screening were dichotomized for each subtest with the exception of Subtest 2 to be either a Weakness to Bolster (greater than 1 standard deviation below the mean) or Not a Weakness (all other scores), and also dichotomized as a Strength to Celebrate (greater than 1 standard deviation above the mean) or Not a Strength (all other scores). It should be noted that Subtest 1 (Language Processing) and Subtest 6 (Calculation) were revised. Six items were added to Subtest 6 (Calculation) in the middle of the second year of data collection. Five items were added to Subtest 1 (Language Processing) and three items removed at the beginning of the third year of data collection. Data from both the original and revised versions for Subtests 1 and 6 were included in the analysis. The gold standard tests did not include a digit span component that could be compared to performance on Subtest 2.

The results of the KDI-II, KBIT-2, and K-SEALS tests were also dichotomized in the same manner. The dichotomized variables were then compared in 2 x 2 contingency tables and Negative Predictive Powers (NPPs) were calculated. The NPP is the probability that a subject with a negative screening test truly does not have a weakness as measured by another instrument. These values should all be above 0.85 to show the expected agreement between the Well Screening and the other standardized tests. The values ranged from 0.85 (Strength of Subtest 3) to 1.0 for multiple tests.

Table 8.9. Contingency table Well Screening® x-KDI-II, KBIT-2, K-SEALS

Fall			**NPP**
Contingency Table			
Subtest 1 (Original)	Weakness to Bolster	KDI-II	0.98
Subtest 1 (Original)	Strength to Celebrate	KDI-II	0.85
Subtest 1 (Revised)	Weakness to Bolster	KDI-II	1.00
Subtest 1 (Revised)	Strength to Celebrate	KDI-II	1.00
Subtest 3	Weakness to Bolster	KDI-II	0.85
Subtest 3	Strength to Celebrate	KDI-II	0.90
Subtest 4	Weakness to Bolster	K-SEALS	0.99
Subtest 4	Strength to Celebrate	K-SEALS	0.89
Subtest 5	Weakness to Bolster	KBIT-2	1.00
Subtest 5	Strength to Celebrate	KBIT-2	0.89
Subtest 6 (Original)	Weakness to Bolster	KDI-II	0.95
Subtest 6 (Original)	Strength to Celebrate	KDI-II	0.97
Subtest 6 (Revised)	Weakness to Bolster	KDI-II	1.00
Subtest 6 (Revised)	Strength to Celebrate	KDI-II	0.93
Subtest 7	Weakness to Bolster	KBIT	1.00
Subtest 7	Strength to Celebrate	KBIT	0.89
Subtest 8	Weakness to Bolster	KDI-II	0.91
Subtest 8	Strength to Celebrate	KDI-II	0.92
Subtest 9	Weakness to Bolster	K-SEALS	n/a
Subtest 9	Strength to Celebrate	K-SEALS	0.90
Subtest 10	Weakness to Bolster	K-SEALS	n/a
Subtest 10	Strength to Celebrate	K-SEALS	0.95

1 = Language Processing; 3 = Word Sound Play; 4 = Confrontational Naming; 5 = Pragmatics; 6 = Calculation; 7 = Language Formulation (morphology); 8 = Letter Recognition; 9 = Real Word Reading; 10 = Nonsense Word Reading; n/a = not applicable because children are not expected to read at the beginning of kindergarten.

SENSITIVITY AND SPECIFICITY

Sensitivity and specificity are needed to fully understand a test's strengths as well as a test's limitations. Sensitivity measures how often a test correctly generates a positive result for individuals who have the condition that is being tested. This is also known as the true positive rate. Specificity measures a test's ability to correctly generate a negative result for individuals who do not have the condition that is being tested. This is also known as the true negative rate.

Data for the sensitivity and specificity analyses were available for 188 of the 274 kindergarten subjects. Subtest 7 (Language Formulation) had a slightly smaller sample size of 188 because it was added to the test battery at the end of the first year of data collection. Students were labeled "at risk" if they failed three or more subtests and "no risk" if they failed fewer than three subtests upon entering kindergarten (fall). A failed score on a subtest was set at −1.00 or more standard deviations below the mean. The actual condition of the student was determined by failure of three or more subtests when tested at the end of the school year (spring).

In Table 8.10, the two columns show the true condition of the student: at risk or no risk. The rows indicate the results of the screener: positive or negative. Cell A (True Positive) includes students identified as at risk when tested in fall and still identified as at risk in spring. Cell D (True Negative) includes students identified with no risk when tested in fall and still identified as no risk in spring. Cell B (False Positive) includes students identified as at risk when tested in fall but no longer identified as at risk in spring. Cell C (False Negative) includes students who were not identified as at risk in fall, but were then identified as at risk in spring.

The prevalence of students identified as being at risk was 13.8%. The sensitivity that the screener correctly identifies children at risk among children who are truly at risk is 61.5%. This lower percent is reasonable given that exposure to academic skills when entering kindergarten is variable. There are children who do *not* perform well when entering kindergarten but who do well after being exposed to skills that were not readily available to them at home or in their preschool experience. It is important to use the Well Screening as a baseline measure in fall and a monitoring tool throughout the school year to measure growth and effectiveness of interventions for all of the children identified as at risk at the beginning of the school year. The specificity that the screener correctly identifies children with no risk among children who are truly at no risk is 96.7%. The probability that the student is truly at risk when tested positive in fall is 72.7%. The probability that the student is not at risk when tested negative in the fall is 94.0%.

Table 8.10. Prevalence, sensitivity, specificity, positive predictive value, negative predictive value

	At risk (number)	**No risk (number)**	**Total (number)**
Positive (number)	**A** True positive 16	**B** False positive 6	Total positive 22
Negative (number)	**C** False negative 10	**D** True negative 156	Total negative 166
	Total at risk 26	Total no risk 162	Total 188

There are many benefits for completing a kindergarten screening with little or no drawbacks. Most important, children who are at risk for language and learning disabilities can be identified for further evaluation, which, in turn, allows for early intervention. Furthermore, children who lag behind their peers because of lack of exposure to academic skills can be identified early, allowing for targeted intervention and monitoring growth. Without completing a screening, it may take classroom teachers or professionals weeks to understand each child's strengths and needs. The Well Screening can help ensure that children's unique learning profiles are quickly identified to ensure that follow-up steps can be taken as needed.