INTERNATIONAL COMMISSION
ON HEALTH CARE CERTIFICATION

Certified Life Care Planner Examination Validation Process and Statistics Report

International Commission on Health Care Certification
13801 Village Mill Drive • Suite 103
Midlothian, Virginia 23114
P: (804) 378-7273 • F: (804) 378-7267
www.ichcc.org

Test Statistics, Test Validation, and Cutoff Test Score Analysis

The ICHCC administers the Certified Life Care Planner examination online with proctor options that include online proctoring through ProctorU.com, a local community college, junior college, college, university, Sylvan Learning Center, or public library. The examination is a timed test that allows 2 minutes for each test item at 100 items, for a total test time of 3 hours and 20 minutes.

All tests are scored by the online test software program at the submission of the last item by the candidate, and the results are sent directly to the corporate office of the International Commission on Health Care Certification. The CLCP examination's cutoff score and item validation were derived and achieved using the Angoff Method (Modified) (Arrasmith and Hambleton, 1988; Ashby, 2001; Biddle, 1993; Bowers and Roby, 1989; Carlson and Strip, 2009; Tiratira, 2009). The ICHCC Test Committee met on June 2-3, 2012, and one of the activities in which 18 Test Committee members participated was the determination of the cutoff test score for the CLCP examination using the criterion-referenced model. The specific model used was the modified Angoff method in which rating participants discussed the characteristics of a borderline certification candidate, and a consensus was reached as to the specific characteristics to consider when reviewing each individual item. The raters were asked, "Would a borderline candidate be able to answer the item correctly?" The items that the Committee felt would be answered correctly by the borderline certification candidate were assigned a 1=yes. Items that the Committee felt that the borderline candidate would more than likely mark a wrong answer were assigned a 0=no. A second meeting of the Test Committee was held on March 1 — 2, 2013, and all items were reviewed and rated a second time by 5 committee members. A total of 208 examinations administered in 2011 through March of 2012 were used in the validation and cut-score determination study. The rater reliability coefficients, Cronbach alpha for internal consistency of ratings, and the cut-score by 3 levels from the 2nd Test Committee ratings are presented in the Tables below.

Table 1— Reliability Coefficients Between Individual Raters

	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1.00			
Rater 2	0.28	1.00		
Rater 3	0.50	0.32	1.00	
Rater 4	0.24	0.18	0.23	1.00
Rater 5	0.19	0.29	0.03	0.13

Discussion: The above matrix displays the correlations between the 5 raters on the item ratings. The positive numbers indicate various levels of "agreement" between raters; low correlations are between .10 and .30; high correlations are .30 to .50; and very high correlations exceed .50. The above display represents a mix of low correlations and high correlations between raters. Rater 5 had the lowest correlations overall.

Table 2 — Overall Reliability by Rater
(Each Rater's Reliability to the Average Rating of All Other Raters)

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Correlation	0.49	0.41	0.45	0.29	0.23
P-value	0.00	0.00	0.00	0.00	0.02

Discussion: This output shows the correlations for each rater indicating how consistent their ratings were relative to all other raters on the panel. All raters are statistically significantly correlated with the average rating of all other raters (with p-values less than .05 significance). Any rater above .05 significance, according to protocols, is removed by the program from the calculation of the overall Critical Score at this step in the process.

Table 3 — Overall Rater Panel Reliability
R = 0.61 Intra-Class Correlation Coefficient

Discussion: this output shows the overall reliability of all raters using the "intraclass correlation coefficient" (ICC), which shows the average reliability to the entire panel as a whole. It is desirable to have a panel with an ICC value that exceeds .50, but lower values may be accepted.

Table 4 - Outlier Raters (Raters With Overall Averages That are Significantly High or Low Compared to the Average of the Panel)

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Average Rating	78.66	81.75	85.85	82.63	82.82
Non-Outliers	78.66	81.75	85.85	82.63	82.82

Discussion: These data highlight raters who, on average, rated items significantly higher or lower than other raters (using a rule of +/- 1.645 standard deviations from the average of the overall panel). The data suggest that the individual raters are within "normal ranges" for their rating averages. The data from raters who are significantly higher or lower than other raters' data are eliminated from the analysis. This process eliminates atypical data reported by subject mater experts participating in the rating process and brings the average values more within "normal ranges" rather than skewing the average based on extreme data points.

Table 5 — Overall Critical Score
With all Raters Included = **82.34**

Discussion: This table presents the overall critical score that represents the final, unmodified Critical Score for the test that will later be reduced by one, two, or thee Conditional Standard Error of Measurements to establish the final cutoff for the test. All raters are included in this final analysis as there were no significant outlier rater data points among raters.

Table 6 Overall Critical Score Results	
Correlation Between Angoff Ratings and Item Difficulty values	**0.28**
Difference between Critical Score and Test Difficulty	**1%**
Skew of Difference Values	**1.30**
Standard Error of Skew	**0.24**
Standard Error of Skew Threshold (2X Standard Error of Skew)	**0.49**
Skewness Test Result (Skew/Standard Error of Skew	**5.31**
Adjusted Critical Scores	
Optimum Critical Score #1	**78.57**
Optimum Critical Score #2	**79.26**

Discussion: The above items in Table 6 are delineated as follows:

1. **Correlation between Angoff Ratings and Item Difficulty Values:** This item provides the correlation between the minimum passing score estimates (Angoff ratings) provided by the

Subject-Matter Experts (the 5 panel raters also referred to as SMEs) and the Item Difficulty Values (also called "item p-values," or the percentage of test-takers who answered the item correctly) from the test-takers. Stronger correlations suggest a tighter connection between the competency levels judged by the SMEs who rated the items and the test-taker pool taking the test. Typically, correlation values fall in the .20s, but correlation values may range between .15 to as high as .55. As noted in Table 6, the correlation value is within an acceptable range at .28 correlation.

2. **Difference between Critical Score and Test Difficulty:** This item provides the average difference between the Angoff ratings (From SME raters) and the Item difficulty Values (from test-takers). Positive values identified in this item indicate that the Angoff ratings were higher than the Item Difficulty Values. For example, if an item had an Angoff rating of 80% (as determined by the raters, noting that they felt 80% of the candidates would answer this item correctly) and 75% of the test-takers answered this item correctly, a 5% difference would be displayed for this particular item (80% - 75% = 5%). As noted in Table 6 under this item, the average difference between all items the raters thought would be answered correctly and those items that were actually answered correctly by the test-takers is 1%, which is a very accurate rater estimate for their projection of the number of items that would be answered correctly by certification candidates testing for the first time.

3. **Skew of Difference Values:** The term "Skew" reflects whether the distribution of data is symmetrical (i.e., uniformly distributed with an equal number of values above and below the average of the distribution). Thus, if the skewness statistic is zero (0), the data are perfectly symmetrical. Regarding the applications of skewness to the Angoff rating method, if the skewness statistic is less than -1 or greater than +1, the distribution of highly skewed. If the distribution is between -1 and -1/2 or between +1/2 and +1, the distribution is moderately skewed. If skewness is between -1/2 and +1/2, the distribution is approximately symmetric. The skew statistic is calculated by obtaining the difference between the Angof£ ratings and the Item Difficulty Values. Positive skew values reveal that there is a disproportionately high number of test items with positive values (i.e., items that were potentially over-rated by the raters). The Skew of Different Values raw statistic for the CLCP items is 1.30, which indicates a tendency for the raters to overrate the CLCP items.

4. **Standard Error of Skew:** The value of the standard error of skew should be close to zero for data to follow a normal distribution. The formula for the standard error of skew is 6/n where n=sample size. If the standard error of skewness is more than twice the standard error of measurement, then the data are positively skewed.

5. **Standard Error of Skew Threshold (2X Standard Error of Skew):** This item determines if the data are skewed either positively or negatively from the test mean. For example, twice the Std. Error of Skewness is 2 X .245 = .49. Setting the skewness threshold range between -.49 and +.49, it is determined if the value for Skew of Difference Values falls within this range. The Skew of Difference Values is 1.30, and is beyond the skew range, suggesting that the data are positively skewed, and thus the data form a non-normal distribution of data.

6. **Skewness Test Result (Skew/Standard Error of Skew:** This item is used to determine whether the skewness of the distribution is significant. The Angoff method identifies high differences between Angoff ratings and Item Difficult Values at 2.0 and greater. If the Skewness Test exceeds 2.0, it is recommended that the OPT Critical Score #1 be used as the cut-score. This score is computed by reducing each over-rated item's Angoff rating to the outer lower limit (1.96 X Standard Error of the Mean of the SME ratings for each over-rated item). If the Skewness Test results exceed 3.0 (5.31 for the SME ratings for this examination), the OPT

Critical Score #2 is recommended. This score is calculated by reducing the Critical Score to the outer lower limit of the raters (1.96 Standard Errors of Difference from the Critical Score, using the average rater reliability and Standard Deviation of the raters' average ratings). The OPT Critical Score #2 provides a greater correction than OPT Critical Score #1. Therefore, the cut-score for the CLCP examination is established at 79, down 3 points from the raw critical score of 82.

The test statistics are presented below. The Test Statistics by Score table illustrates the Mean, Standard Deviation, the Standard Error of Measurement, and the minimum and maximum scores.

Table 7 — Test Statistics by Score

Mean	Standard Deviation	Standard Error of Minimum Measurement		Maximum
80.236	7.511	3.317	24.000	100.000

Test reliability is presented with three reliability estimates that include Cronbach's Alpha, Guttman's Split Half, and KR-21. Cronbach's Alpha is a widely accepted method for determining the **internal consistency** of a written test. The reliability using this method is shown, along with interpretive guidelines of Excellent, Good, Adequate, and Limited, which are taken from the U.S. Department of Labor's guidelines (DOL, 2000). The Guttman split-half reliability coefficient is an adaptation of the Spearman-Brown coefficient, but one that does not require equal variances between the two split forms. The KR-21 formula is another method for evaluating the overall consistency of the test. It is typically more conservative than the Cronbach's alpha, and is calculated by considering only each applicant's total score, whereas the Cronbach's Alpha method takes item-level data into consideration. The reliability coefficients are illustrated in Table 8.

Table 8 — Test Reliability

	Cronbach's Alpha	Guttman Split Half	KR-21
Coefficient	0.805	0.817	0.726
Quality Rating	Good	Good	Adequate

The modified Angoff process determines the critical point in the score distribution that delineates "qualified" from "unqualified" based on Subject Matter Expert (SME) ratings, the measurement properties of the test, and the consistency and accuracy of the SMEs. There are two steps involved in the administration of the modified Angoff process; 1) a panel of raters determines the Critical Score, which is the average of their Angoff ratings for all of the items included on the test, and 2) reduce the Critical Score using one, two, or three Conditional Standard Error of measurements (CSEM) (to account for the measurement error of the test), which provides three cutoff options for the test.

The Standard Error of measurement (SEM) of the test is calculated by multiplying the standard deviation by the square root of 1 minus the overall test reliability. The SME formula uses Cronbach's Alpha for the calculation. The SEM provides a confidence interval of an applicant's true score around his or her obtained score. An applicant's *true score* represents his or her true, actual ability level on the overall test; whereas an applicant's *obtained score* represents where he/she just happened to score on that given test day. For example, if the test's SEM is 3.0 and an applicant *obtained* a raw score of 60, his or her true score (with 68% likelihood) is between 57 and 63, between 54 and 66 (with 95% likelihood), and between 51 and 69 (with 99% likelihood).

The preferred Standard Error of Measurement (SEM) is the *Conditional* Standard Error of Measurement (CSEM) when setting cutoff scores as opposed to the traditional Standard Error of Measurement (Standards, 1999). The CSEM provides an estimate of the SEM for each score in the distribution, allowing the user to focus on the CSEM in the range of scores around the Critical Score, which is the area of decision-making interest (Standards, 1999). The classical SEM provides only an average that considers all scores in the distribution. Because the SEM considers the average reliability of scores throughout the entire range of scores, it is less precise when considering the scores of a particular section of the score distribution.

Cutoff Options/Adverse Impact: This output incorporates three cutoff scores and the Decision Consistency Reliability and Kappa Coefficient. These data for the proposed cutoff scores are illustrated in Table 9.

Table 9 — Final Cutoff Options/Adverse Impact

			Interpretation for Mastery-based Tests					
			Decision Consistenc		i Reliability	Kappa Coefficient		
Cutoff Options	Cutoff Score		Estimated	Calculated	Interpretation	Estimated	Calculated	Interpretation
	Raw Points	Percent						
A	85	85.00%	0.82	0.80	Adequate	0.58	0.50	Good
B	81	81.00%	0.80	0.76	Adequate	0.59	0.52	Good
C	78	78.00%	0.80	0.77	Adequate	0.59	0.51	Good

Discussion: The table items are clarified as enumerated below:

1. **Cutoff Score:** The Program automatically populates this field with the Critical Score (unmodified Angoff percentage score from SMEs) based upon the "final" percentage score, or the Critical Score as presented in Table 5.

2. **Decision Consistency Reliability:** the DCR is the appropriate type of reliability to consider when interpreting reliability and cutoff score effectiveness for mastery-based tests. Mastery-based tests are tests used to classify examinees as "having enough competencies" or "not having enough competencies" with respect to the Knowledge, Skills Abilities, and Personal Characteristics (KSAPC) set being measured by the test. The DCR attempts to answer the following questions regarding competency-level cutoff on a test: If the test was hypothetically administered to the same group of examinees a second time, how consistently would the test pass the examinees (i.e., classify them as masters) who passed the first administration *again on a second administration?* Similarly, DCR answers: How consistently would examinees who were classified by the test as non-masters (failing) fail the test the second time? This type of reliability is different than internal consistency reliability (e.g., Cronbach's Alpha and KR-21), which considers the consistency of the test internally, without respect to the consistency with which the test's cutoff classifies examinees as masters and non-masters.

3. **Kappa Coefficient:** A Kappa coefficient explains how consistently the test classifies masters and non-masters beyond what could be expected by chance. This is essentially a measure of utility for the test. Kappa coefficients exceeding .31 indicate adequate levels of effectiveness and levels of .42 and higher are good.

The summary of test results by gender and the passing percentages for all three cutoff scores are presented in Tables 10 and 11 below.

Table 10 — Summary Test Results by Gender

Total Number	Mean	Standard Deviation	Standard Mean Group Difference
Total Test Takers (208)	80.236	7.511	N/A
Men (40)	79.800	6.014	N/A
Women (168)	80.339	7.837	-0.072

Table 11 — Number Passing at Each Cutoff

	Cutoff A	Cutoff B	Cutoff C
All Test Takers (208)	54 - 26%	109 - 52%	158 — 76%
Men (40)	9 — 23%	19 — 48%	28 — 70%
Women (168)	45 - 27%	90 - 54%	130 - 77%