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 Certified Life Care Planner was validated and its cut-score calculated using the Biddle Test Validation and Analysis Program (2010). This program uses a content validity approach for validation and is designed for use by human resource professionals to aid in validating and analyzing written tests. It is most appropriate for validating tests designed to measure knowledge, skills, abilities, and personal characteristics ("KSAPCs"). The TVAP program relies on two workbooks for entering and analyzing data: the **Test Validation Workbook** and the **Test Analysis Workbook**. When used in sequence, these workbooks provide a complete set of tools to validate written test (entry level or promotional), analyze item-level results (after administration), and set job-related and defensible cutoff scores using court approved methods. The Biddle program utilizes the job analysis of life care planning for the Certified Life Care Planner (CLCP) credential and a draft of the written test that is based on the job analysis (role and function study).

The Test Validation Workbook is used in validating the examination before the test is offered to the CLCP candidates. A survey is included in the workbook that a group of Subject Matter Experts (SMEs) use to rate each item on the examination being reviewed (test items). The SMEs enter their ratings into the (Test Validation Workbook) and analyzed to evaluate which items to include on the final test and which items to discard, or save for re-evaluation by the SMEs after revision. After the final items are identified using the Workbook, part of the SME ratings (the Angoff ratings that reflect the opinions from SMEs regarding the minimum passing level for each item on the test) are exported to the Test Analysis Workbook for the remaining steps in the process.

The Test Analysis Workbook is used to analyze the test results (at both an item and overall test level), modify and improve the test based on these results, and then set job-related cutoffs using methods that have been previously endorsed by the courts (Contreras v. City of Los Angeles and U.S. v. South Carolina) (See Appendix A). These court cases provided guidelines for establishing minimum thresholds (71% and 63% respectively) for the levels of SME endorsement necessary for screening test items for inclusion on a final test. The Biddle TVAP program uses a ">65% job duty or KSAPC linkage" criteria for classifying items as acceptable for inclusion on a test. If less than 65% of the SMEs link the item to a job duty or KSAPC, the program "red-flags" the item for a closer evaluation.

The ICHCC Test Committee SME members met on June 2-3, 2012, and one of the activities in which 18 SMEs 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 SME members. The Biddle TVAP program was used and 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 the degree of "agreement" between raters. No negative reliability coefficients were obtained, which would indicate "disagreement" in item ratings. Based upon the conventions set by Cohen (1988) ; 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. The correlation between Rater 3 and Rater 5 demonstrates the least amount of agreement (while still positive) between two raters.

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). Any rater who attained a reliability coefficient with a p-value above .05, according to protocols, would have been removed by the program from the calculation of the overall Critical Score at this step in the process. Based upon these outcomes, all raters and their ratings were included in the calculation of the overall critical score.

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: The Outlier Raters table provides the mean Angoff rating for each rater. 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) would be removed from the calculation of the overall critical score. 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. The data suggest that the individual raters are within "normal ranges" for their rating averages. No raters from this study fell outside the "normal range" and all Angoff ratings were used in the calculation of the overall critical score.

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 values in Table 6 are calculated based upon the responses of 208 test takers who were presented with the actual test. Of the 208 test takers, 40 were men and 168 were women. 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 Angoff 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 $\sqrt{6/n}$ 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.

7. **Adjusted Critical Scores**: 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 Measurement	Minimum	Maximum
80.236	7.511	3.317	24.000	99.000

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. The cutoff options listed below offer cutscores that can be used to reduce adverse impact in the event adverse impact is found at the Optimal Cutoff Score #2 (79).

Table 9 – Final Cutoff Options/Adverse Impact								
			Interpretation for Mastery-based Tests					
			Decision Consistency Reliability			Kappa Coefficient		
Cutoff Options	Cutoff Score		Estimated	Calculated	Interpretation	Estimated	Calculated	Interpretation
	Raw Points	Percent						
A	76	76.00%	0.83	0.80	Adequate	0.57	0.50	Good
B	72	72.00%	0.89	0.88	Good	0.53	0.45	Good
C	68	68.00%	0.95	0.95	Excellent	0.47	0.38	Adequate

Discussion: The table items are clarified as enumerated below:

1. **Cutoff Score**: Each cutoff option is one additional conditional standard error of measurement (CSEM) below the Optimum Critical Cutscore #2 of 79. Lowering the cutoff score by a CSEM may be used to reduce potential adverse impact by allowing additional test taker to pass, at the risk of reducing the test's utility.

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)	175 – 84%	199 – 96%	205 – 99%
Men (40)	32 – 80%	38 – 95%	39 – 98%
Women (168)	143 – 85%	161 – 96%	166 – 99%

Annual Reviews

The key to ensuring acceptable validity of the items and that they retain a linkage to the Pomeranz, Yu, and Reid (2010) role and function study (job/practice analysis) is to hold annual test-committee workshops regarding the rating of new items as they are written and applied to the certification examination. The ICHCC conducted two separate Test-Committee workshops prior to the final rating of the items by the subject matter experts (SME) in 2013 as noted above. The item validation workshop and the items' reliability towards the Pomeranz (2010) role and function study will continue on an annual basis as new items are developed. This workshop is scheduled for October of each subsequent year.

CLCP Population Test Statistics

The ICHCC began testing health care professionals for the Certified Life Care Planner credential in March of 1996. A statistical review of scores and years experience among the health care professionals who have eranred the CLCP credential are illustrated in Table 11.

Table 11 - CLCP Descriptive Statistics for Years Experience and Scores

	Years Experience	Scores
N of Cases	1309	1320
Minimum	1.000	0.000
Maximum	45.000	100.00
Mean	18.585	77.086
Standard Deviation	8.64	8.763

Discussion: The total number of Certified Life Care Planners is 1,320. The minimum number suggests a "0" minimum score on the examination, but is actually due to a test-case score when conducting a trial run of the test-online software program. Please note the difference in the N of Cases between Years Experience and Scores. There are 11 cases that did not report their years of experience for beginning their careers post-award of their formal degrees to the present, or at the time they completed their CLCP test applications.

It is important for the ICHCC to determine if one's years-of-experience influenced the test scores among the specialty health care fields represented within the 1,320 health care service providers. The Years-of-Experience statistical results are presented in Table 12.

Table 12 - Analysis of Variance on Years of Experience Influence on Scores

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Years	5741.696	66	86.995	1.138	0.215
Error	94332.495	1234	76.444		

$p < .05$

Discussion: A One-Way Analysis of Variance (ANOVA was the chosen statistic in which test scores were set as the dependent variable and Years-of-Experience was used for the independent variable. The Null Hypothesis was stated as: H_0: There are no significant differences among scores of CLCP candidates who have more years of experience vs. test scores of those CLCP candidates with lesser years of experience. The alpha level was set a .95, where .05 is the level of significance for this study. The F-ratio was 1.138 which exceeds the .05 level of significance, thus accepting the null hypothesis that years-of-experience has no significant influence on scores.

There have been a varying number of professions represented within the total number of persons who successfully passed the CLCP examination. It is important for the ICHCC to determine if any of the professions represented in the 1,320 CLCPs had higher group scores than the other heath care related

professions. Such information will help the ICHCC identify any possible item-bias regarding emphasizing any one professional group's training and expertise over all other professional groups represented in the development of its test items. The professions are identified as follows:

1. ARN = Associate Degree in Nursing
2. BSN= Bachelor's Degree in Nursing
3. MSN = Master's Degree in Nursing
4. BABS = Bachelor's of Arts and Bachelor's of Science in Health Related Professions Other than Nursing
5. MAMS = Master's of Arts and Master's of Science in Health Related Professions other than Nursing
6. Ph.D. = Doctorate Level Degree(to include Ed.D., Psy.D, Rh.D.) in Health Related Professions other than Nursing
7. MDDO = Medical Doctor and Doctor of Osteopathic Medicine
8. DC = Chiropractor

Multiple Regression Analysis of Variance of Formal Degree Influence on Scores					
Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
ARN	539.988	1	539.988	7.133	***0.008**
BSN	168.163	1	168.163	2.221	0.136
MSN	10.963	1	10.963	0.145	0.704
BABS	2.156	1	2.156	0.028	0.866
MAMS	75.053	1	75.053	0.991	0.320
PHD	276.538	1	276.538	3.653	0.056
MDDO	605.332	1	605.332	7.996	***0.005**
DC	55.834	1	55.834	0.738	0.391
Error	98569.493	1302	75.706		

***p > .05**

Discussion: A multiple regression ANOVA model was used to determine if any of the formal degreed fields held by the 1,320 certified life care planners influenced the CLCP examination scores significantly over any of the other degreed fields represented. The null hypothesis is stated as: H_0 There are no differences in score levels among the formal degree fields represented by the 1,320 persons who successfully passed the examination. The results showed that the Associate Degree Nurses and the Medical/Osteopathic doctors had mean scores that deviated from the population means significantly, suggesting that their scores were significantly higher than the other degree fields represented. These data results replicate an earlier multiple regression analysis of degree-group scores in 2004 in which medical/osteopathic doctors and associate degree/diploma nurses scored higher on the examination than all of the other represented degreed groups. Thus, the null hypothesis is rejected.

References

Cohen, J. (1988). ***Statistical power analysis for the behavioral sciences*** (2nd ed.). Hillsdale, NJ: Erlbaum.

Pomeranz, J., Yu, and Reid, C. (2010). Role and function study of life care planners. ***Journal of Life Care Planning***, 9(3), 57-106.

Standards for educational and psychological testing. (1999). American Educational Research Association: Washington, D.C.

TVAP: test validation & analysis program. (2010). Version 7.0. Biddle Consulting Group, 193 Blue Ravine, suite 270, Folsom, CA.

U.S. Department of Labor: employment and training administration (2000). ***Testing and assessment: an employer's guide to good practices***.

APPENDIX A

Biddle Court Cases

Contreras v. City of Los Angeles: This case involved a three-phase process used to develop and validate an examination for an Auditor position. In the final validation phase, where SMEs were asked to identify a knowledge, skill, or ability that was measured by the test item, a "5 out of 7" rule (71%) was used to screen items for inclusion on the final test. After extensive litigation, the Ninth Circuit approved the validation process of constructing a written test using items that had been linked to the knowledge, skills, abilities, and personal characteristics (KSAPCs) of a job analysis by at least five members of a seven-member SME panel.

U.S. v. South Carolina: SMEs were convened into ten-member panels and asked to provide certain judgements to evaluate whether each question on the tests (which included 19 subtests of a National Teacher Exam used in the state) involved subject matter that was a part of the curriculum at his or her teacher training institution, and therefore appropriate for testing. These panels determined that between 63% and 98% of the items on the various tests were content valid and relevant for use in South Carolina. The U.S. Supreme court endorsed this process as "sufficiently valid" to be upheld.

Discussion: These cases provide guidelines for establishing minimum thresholds (71% and 63% respectively) for the levels of SME endorsement necessary for screening test items for inclusion on a final test to be used for selection or promotion purposes. In either case, it is important to note that at least an "obvious majority" of the SMEs was required to justify that the items were sufficiently related to the job to be selected for inclusion on the test.

Following the reasonable precedence established by these two cases, this Program uses "> 65% job duty or KSAPC linkage" criteria for classifying items as :acceptable for inclusion on a test. If less than 65% of the SMEs link the item to a job duty or KSAPC, the Program simply "red flags" the item for a closer evaluation. It should be noted, however, that because this Program function requires at least a 65% endorsement level for each item, the collective endorsement level for all of the items used for an actual test is likely to be much higher. This is because several of the items are likely to have much higher SME endorsement levels.