Practical Assessment,
Research & Evaluation

A peer-reviewed electronic journal.

Volume 29 Number 6, March 2024 ISSN 1531-7714

A Practical Comparison of Decision Consistency Estimates

Amanda A. Wolkowitz (*Data Recognition Corporation*), Russell Smith (*Alpine Testing Solutions, Inc.*)

A decision consistency (DC) index is an estimate of the consistency of a classification decision on an exam. More specifically, DC estimates the percentage of examinees that would have the same classification decision on an exam if they were to retake the same or a parallel form of the exam again without memory of taking the exam the first time. This study compares three classical test theory DC estimates in the context of high stakes pass/fail exams. The three methods compared include those developed by Livingston and Lewis (1995), Peng and Subkoviak (1980), and Wolkowitz (2021). This study compares the computationally and conceptually simpler DC methods proposed by Peng-Subkoviak and Wolkowitz to the more widely used and accepted, but more complex, method proposed by Livingston and Lewis. Through a comparison of two simulated datasets and three operational datasets, the results suggest that the Livingston-Lewis and Wolkowitz methods produce relatively similar results for datasets with skewed distributions and all three methods produce reasonably similar results for normally distributed datasets. Following these results, this study provides guidelines for deciding which method to apply as well as industry guidelines for acceptable DC values.

Keywords: decision consistency, reliability

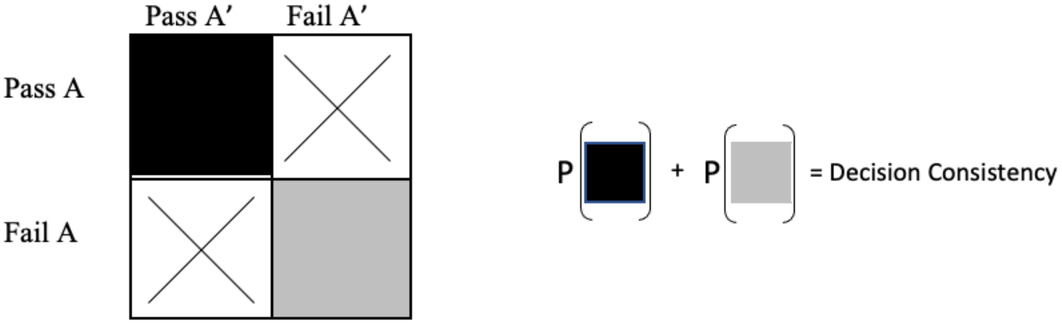
Introduction

Decision consistency (DC) is a measure of reliability that estimates the proportion of examinees who are classified into the same category on two administrations of parallel forms of an exam. In professional credentialing, there are typically two classifications: pass or fail. The estimation of the true DC index requires two administrations of parallel forms of an exam. Because it is not possible to have examinees take an exam and then retake the same or parallel form of the exam under identical conditions without remembering the first experience, DC is often estimated using an observed distribution and a predicted or true distribution.

DC is used in the industry to determine "the extent to which the observed classifications of examinees would be the same across replications of the testing procedure" (AERA, APA, & NCME, 2014, p. 40). Since the goal is to have as consistent of a pass/fail decision as possible in professional certification and licensing, the value of this statistic is important and should be evaluated on such exams. The guideline provided in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) states: "When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure (Standard 2.16, p. 46)." The method used to estimate the DC index is one of choice.

Figure 1 illustrates the concept behind DC. Form A is the form of the exam that examinees complete. Form A´ is a hypothetical equated, parallel form of the same exam that all examinees "complete". The black box represents a consistent pass decision. The grey box represents a consistent fail decision. The "X" boxes represent inconsistent decisions. DC is the sum of the probabilities of the consistent decisions.

Figure 1. Probability of Passing and Failing Two Attempts at a Parallel Set of Exam Forms



DC can be more formally defined as the probability that examinee *i* will obtain the same classification decision on equated, parallel Forms A and A´. As illustrated in Figure 1, there are two ways in which an examinee may obtain a consistent decision on a pass/fail exam: pass on both forms or fail on both forms. This can be written as follows, where X = scores on Form A, X´ = scores on Form A´, and C = criterion score (or passing/cut score):

Probability of examinee *i* passing both forms =

$P(X_i \geq C) \cdot P(X'_i \geq C)$

Probability of examinee *i* failing both forms =

$P(X_i < C) \cdot P(X'_i < C)$

Probability of examinee *i* having a consistent decision:

$P_{o(i)} = P(X_i \geq C) \cdot P(X'_i \geq C) + P(X_i < C) \cdot P(X'_i < C)$ (1)

As stated above, the probability of a consistent decision is the sum of the consistent decisions for all examinees:

$$P_o = \frac{\sum_{i=1}^{N} P_{o(i)}}{N} \qquad (2)$$

It may sometimes be useful to think of the probability of a consistent decision as the inverse of an inconsistent decision, P_x, or simply $P_o = 1 - P_x$.

Cohen (1960) introduced an adjustment to the DC estimate provided in Equation 2 to account for chance. This is shown in Equation 3, where P_e is the expected agreement of the pass/fail decision that would occur by chance.

$$\varkappa = \frac{P_o - P_e}{1 - P_e}, \qquad (3)$$

Cohen's kappa was intended to provide a coefficient of *nominal* scale agreement between two judges, i.e., an interrater reliability. On pass/fail written exams, the agreement estimate is not about the consistency of two judges rating an examinee, but about the consistency of the pass/fail decision for one examinee on two different exam administrations.

Aside from the difference in the intended application of the statistic, Cohen's kappa also has an associated paradox that high levels of agreement may lead to low values for kappa. This paradox stems from the use of marginal sums in the computation (Feinstein & Cicchetti, 1990; Cicchetti & Feinstein, 1990; Gwet, 2002). For example, consider 100 examinees who complete two attempts at the same form of an exam and the second attempt is completed without memory of the first attempt. Of those 100, assume 90 have the same pass decision on both attempts, five have the same fail decision on both attempts, two passed on the first attempt but failed on the second attempt, and three passed on the second attempt but failed on the first attempt. The observed agreement in this situation is 95%. The expected agreement is 86%, i.e., P(passing 1st attempt)·P(passing on 2nd attempt) + P(failing on 1st attempt)·P(failing on 2nd attempt) = 0.92·0.93 + 0.08·0.07 = 0.8612. Cohen's $\varkappa$ yields a consistency statistic of 0.64, i.e., much lower than one may expected given the known agreement. This paradox is a reason that Cohen's kappa is not commonly used (if at all) to calculate DC on a high stakes pass/fail exam, such as a licensure exam. It is also why other methods that have expanded on Cohen's kappa to make it possible to estimate DC using a single test administration instead of two (e.g., Cohen, 1968; Huynh, 1976, Subkoviak, 1976, Marshall & Haertel, 1975) have received criticism (Mellenbergh & van der Linden, 1979).

While Cohen's kappa may not be ideal for estimating DC for pass/fail exams, others have developed estimation methods for this purpose. These methods also introduce ways to estimate the DC index using a single test administration. For example, Subkoviak (1976) presented a method in which a DC value is estimated for each examinee by applying the binomial distribution to an examinee's true probability of a correct item response and then averaging the estimated DC values across all examinees. While this method is not commonly used, it is one of the earliest publications offering a way to accurately measure DC on a pass/fail exam with just one administration of the exam.

While there have been many methods since the 1970s that have provided ways to estimate DC, the Livingston and Lewis (1995) method is one of the most widely used and accepted methods and often referenced in research comparing different DC estimate methods (e.g., Young & Yoon, 1998; Li, 2006; Wan, Brennan, & Lee, 2007; Deng, 2011; Alger, 2016). The normal approximation method (Huynh, 1976) that was improved upon by Peng and Subkoviak (1980) is a computationally simpler method than Livingston and Lewis that is also used in the industry. More recently, Wolkowitz (2021) introduced an even more computationally simple method for estimating DC with just one administration of the exam.

Many of the methods used to estimate DC have been computationally tedious (Breyer & Lewis, 1994). For example, Subkoviak (1988) noted that while methods proposed by Livingston and Lewis (1995) and Subkoviak (1976) estimate the DC index, they require knowledge of specific software and background in test theory to fully understand how the method works. Simpler methods, such as those proposed by Peng and Subkoviak (1980) and Huynh (1976) are simpler than earlier methods, but still require the use of bivariate and univariate normal distributions. For non-measurement professionals, such as classroom educators, these distributions may be less familiar to them and difficult to understand the underlying theory. The Wolkowitz (2021) method is a much simpler method both computationally and conceptually; however, it lacks a strong theoretical background like earlier methods.

The purpose of this study is twofold. First, this study aims to compare the DC estimates of three classical test theory (CTT) methods: Livingston and Lewis (LL-DC), Peng and Subkoviak's Normal Approximation (PS-DC), and Wolkowitz (W-DC). The goal is to determine if the simpler and less complex methods of the PS-DC and W-DC methods produce estimates comparable to that of the more computationally and conceptually complex LL-DC method. While there are many CTT methods that could be used for comparison, including item response theory methods, the authors chose to focus on the LL-DC method because it is widely used and accepted and selected two computationally simpler CTT methods, i.e., PS-DC and W-DC. The reason the authors did not include IRT methods in the comparison is because the goal was to determine if simpler, less computationally and conceptually complex methods produced similar methods to the estimates produced by the LL-DC method. Unless non-measurement professionals, such as classroom educators or program directors, have had some measurement courses, IRT methods would not be simpler than CTT methods.

The second purpose of this study aims to provide guidance for acceptable DC values. It is noteworthy that it is not a purpose of this study to determine which DC method is the most accurate under different circumstances since one cannot measure "true" DC. This is akin to reliability measures of internal consistency [e.g., Cronbach's (1951) alpha or McDonald's omega (1999)] in that there is neither one method that produces a "true" measure of reliability nor one that produces a "true" measure of DC. Instead, users must consider the assumptions of the different methods and apply the one that seems most appropriate for their data.

Livingston and Lewis (1995) Method

The LL-DC method is one of the most widely used methods to estimate the DC index. The LL-DC method has four inputs:

1) distribution of the scores on one form of the test,

2) the reliability coefficient of the scores,

3) the minimum and maximum possible scores for the test, and

4) the cut score.

This method applies the notion of an effective test length. As defined by Livingston and Lewis (1995), the effective test length "is the number of discrete, dichotomously scored, locally independent, equally difficult test items necessary to produce total scores having the same precision as the scores being used to classify the test takers" (p. 180). This definition allows this method to be applied to both dichotomously and polytomously scored exams. This flexibility is one of the reasons why this method is applied in practice.

Livingston and Lewis (1995) describe seven steps for implementing their method. The first step is to estimate the effective test length (see p. 187). Next, estimate the distribution of the proportional true scores from the observed score distribution (see Livingston & Lewis, 1995, pp.182, 188). As originally implemented by Livingston and Lewis, this step assumes that the distribution of the proportional true scores has the form of a four-parameter beta distribution (LL-DC-4); however, it may also be implemented under the assumption that the distribution has the form of a two-parameter distribution (LL-DC-2). Then, estimate the conditional distribution of classifications on a parallel form of the exam for examinees at each true-score level using a binomial distribution with parameters *n* and *p*, where *n* is the number of items and p_i is the probability of examinee *i* correctly responding to that item (see Livingston & Lewis, 1995, pp. 182-4). The last several steps involve estimating the joint distribution of classifications to determine decision accuracy as well as consistency (see Livingston & Lewis, 1995, pp. 184-186).

As just described, the LL-DC method is a computationally complex method and requires software, such as BB-CLASS (Brennan, 2004) or the R package *betafunctions* (Haakstad, 2022). Livingston and Lewis suggest that the LL-DC method will work with different score distributions including extremely skewed data. However, this method has been found to be more sensitive to reliability estimates and score distributions compared to other DC estimates (Wan, Brennan, & Lee, 2007; Deng, 2011) and has larger biases for exams with a small number of items (Li, 2006; Deng, 2011).

For purposes of this study, it is important to note that the binomial distribution assumption is part of the LL-DC method. This distribution assumes that the number of observations is fixed, the observations are independent of one another, each observation is binary (i.e., success of failure), and the probability of success is the same for all items for examinee *i*. The first three assumptions are reasonable assumptions for most professional credentialing exams in which there are no testlets or dependencies within the exam. However, the last assumption is technically violated when analyzing exam data because an examinee does not have an equal chance of correctly responding to each item on an exam. While there is some agreement to the robustness with respect to violations of this assumption in DC estimation methods (Subkoviak, 1976; Wan, Brennan, & Lee, 2007), it is a violation worth noting.

Peng-Subkoviak's (1980) Simple Normal Approximation Method

Peng and Subkoviak's (1980) simple normal approximation method is an extension of Huynh's normal approximation procedure (Huynh, 1976). Similar to the underlying assumption in Livingston and Lewis's method, this method assumes that an examinee has an equal chance of responding to each item on an exam. The extent to which this method is robust to the violation of this assumption impacts the strength of the DC estimate.

The first step in the simple normal approximation procedure is to compute the probability P_1 that a standardized normal variate is less than z, where $z = (c - 0.5 - \mu)/\sigma$ and c = criterion score, 0.5 is a correction factor (see Hays, 1973, p. 309), μ = mean of the score distribution, and σ = standard deviation of the score distribution. Next, a bivariate normal distribution table is used to compute the probability P_2 that two standardized normal variates with reliability α (or KR21) are less than z. Then, substitute P_1 and P_2 into Equation 4:

$$\text{PS-DC} = 1 - 2(P_1 - P_2) \quad (4)$$

Since this procedure uses a normal distribution to estimate probabilities, the effectiveness of the DC estimate may be negatively impacted when the distributions of the data are non-normal.

PS-DC method works well with normally distributed data and, like the LL-DC method, is sensitive to reliability estimates (Wan, Brennan, & Lee, 2007). Thus, while the PS-DC method is computationally simpler than LL-DC method, the potential for more error in non-normal data is a possible reason that it is applied less frequently in the industry.

Wolkowitz (2021) Method

The W-DC method (Wolkowitz, 2021) is a computationally simple method for estimating DC. This method makes no assumptions about normality nor does it use the binomial distribution. Instead, it uses the 95% confidence intervals about observed scores and calculates probabilities of a consistent decision for each observable score. Specifically, the first step to estimating W-DC is to construct a frequency distribution of the total scores, *X*. Then, calculate a 95% confidence interval around each observed score *x* in X. Within each confidence interval centered at score *x*, determine the probability that an examinee with score *x* will have a consistent pass/fail decision. Multiply this probability by the observed number of examinees scoring *x*. This is the estimated number of examinees in the sample who would have the same pass/fail decision on a second administration of the same or parallel form of the exam. Finally, the W-DC index equals the sum of the number of estimated consistent decisions across all possible total scores divided by the total number of examinees in the sample. While this method lacks a theoretical basis of using a 95% confidence interval specifically, the results from Wolkowitz (2021) show that the confidence interval works well from a practical standpoint and produces reasonable results.

Comparison of Methods

The LL-DC and PS-DC methods have been compared in the literature. The W-DC method is new and there is very limited research comparing this method to more established methods. In a study by Wan, Brennan, and Lee (2007), the authors noted that the LL-DC and PS-DC methods have shown to perform similarly when the data is normally distributed. When the data is not normally distributed, the LL-DC method performs better due to its assumption of a beta-binomial distribution versus PS-DC's normality assumption. The authors also noted a disadvantage of the LL-DC method is that it does not consider examinees' original pass/fail status; instead, it only uses marginal distributions of the exam scores. They also stated that both the LL-DC and PS-DC methods have been shown to be sensitive to reliability estimates. In a study by Wolkowitz (2021) in which real and simulated datasets investigated the similarity and accuracy of the W-DC and LL-DC methods across multiple different scenarios (i.e., different score distributions, sample sizes, and different reliabilities of the exam scores), the results indicated that the two methods produced similar results regardless of the situation.

Table 1 compares the LL-DC, PS-DC, and W-DC methods. All three methods require total scores, reliability, and the cut score as input. The LL-DC method has the additional required input of the minimum and maximum possible scores on the exam. The LL-DC method assumes a beta-binomial distribution and makes assumptions regarding the calculation of the conditional error variance and the constant of proportionality used in calculating errors of measurement. Livingston and Lewis (1995) note that this latter assumption is a weakness to the model because the estimates of the conditional standard of error measurement are sensitive to the score range. However, they indicate that the model is fairly robust to violations of this assumption. The main assumption of the PS-DC method is bivariate normality. Thus, this method is likely to perform less well with skewed data. The W-DC method does not make any assumptions about the data but makes a practical, but non-theoretical assumption that a 95% confidence interval is the best interval to use in the application of this method.

Examples

An example dataset is used below to help illustrate the similarities and differences in the three methods. This example contains a hypothetical 10-item exam completed by 100 examinees with an alpha reliability of 0.60 and cut score of 5. Table 2 displays the frequency distribution of this hypothetical exam.

All three methods require the following inputs: total score distribution (only the mean and standard deviation of the total scores is needed for PS-DC method), reliability, and passing score. The simple normal approximation method requires a way to estimate values on the bivariate normal distribution table.

Tables 3-5 display the LL-DC, PS-DC, and W-DC estimates, respectively. The LL-DC estimate was computed using BB-CLASS (Brennan, 2004). The intermediate PS-DC values include $z = 0.471$, $P_1 = 0.681$, and $P_2 = 0.550$. The W-DC estimate was

Table 1. Comparison of Three DC Methods

	LL-DC	PS-DC	W-DC
Inputs	Total scores, reliability, cut score, minimum and maximum possible scores	Total scores, reliability, cut score	Total scores, reliability, cut score
Use of hypothetical datasets?	Yes, generates hypothetical exam scores using the beta-binomial model	Yes, generates hypothetical exam scores using a normal distribution	No, uses observed score data only
Assumptions	1. Beta-binomial distribution 2. The conditional error variance of scores on an *n*-item exam for examinees with a given proportional true score equals the variance of a binomial distribution based on *n* observations with a success probability equal to the proportional true score (Livingston & Lewis, 1995, p. 187) 3. "Errors of measurement are proportional to those that would be generated by a binomial distribution. The constant of proportionality depends on the relationship between the possible score range of the test and its estimated effective test length" (Livingston & Lewis, 1995, p. 189)	Bivariate normality	95% confidence interval around a given cut score is an accurate interval for estimating DC
Impact of violating assumptions	Limited research on this topic, but Livingston-Lewis (1995) indicate the method is robust against violating #3	If observed data is non-normal, DC estimates may be less accurate near the mode of the data	Limited research on this topic, but using intervals other than 95% may affect the accuracy of the DC estimates
Computer program requirements	Software programs, such as BB-CLASS or R	Bivariate normal distribution table; software programs very helpful	None
Potential implementation errors	If cut score lands in a region where there are no examinees, programs such as BB-CLASS and R (betafunctions) fail to run	None observed	None observed
Complexity of understanding how the method works	High	Medium	Low

Table 2. Data from a Hypothetical 10-Item Exam

Score	**Freq**
0	2
1	4
2	5
3	7
4	11
5	15
6	22
7	12
8	13
9	5
10	4
Total	100

Table 3. LL-DC Results Using Hypothetical 10-Item Exam and a 4-parameter beta distribution

	Pass A	Fail A	TOTAL
Pass A´	0.55	0.13	0.68
Fail A´	0.13	0.20	0.33
TOTAL	0.68	0.33	1.00

LL-DC = 0.55 + 0.20 = 0.75

Table 4. PS-DC Results Using Hypothetical 10-Item Exam

	Pass A	Fail A	TOTAL
Pass A´	0.54	0.13	0.68*
Fail A´	0.13	0.19	0.32
TOTAL	0.68*	0.32	1.00

*Total does not appear to equal sum of 0.54 and 0.13 due to rounding.

PS-DC = 0.54 + 0.19 = 0.73

Table 5. W-DC Results Using Hypothetical 10-Item Exam

	Pass A	Fail A	TOTAL
Pass A´	0.61	0.13	0.73
Fail A´	0.10	0.16	0.27
TOTAL	0.71	0.29	1.00

W-DC = 0.61 + 0.16 = 0.77

computed by hand, but replicated using the R-code provided in the appendix.

In comparing Tables 3-5, there are a few notable observations. First, the LL-DC and PS-DC methods produce similar tables. This is not a surprising result since this data is approximately normally distributed, both methods use the binomial distribution, and other studies (e.g., Wan, Brenna, & Lee) have found that these two methods produce similar results. Another observation is that the observed pass rate on this exam is 71%. Since the W-DC method is based only on observed data, Table 3 displays this pass rate in the marginal sum representing the percent passing Form A. The other two methods display approximations of this pass rate. Another observation is that the estimated proportion of examinees who pass Form A but fail Form A´ equals the proportion of examinees who fail Form A, but pass Form A´ in the PS-DC and

LL-DC methods. This happens because the PS-DC method uses a normal distribution to estimate the percent of examinees below the standardized score z and this approximation is the same for both forms of the exam. The LL-DC method also makes use of symmetry in the distribution. The W-DC method does not show equal proportions because this method only uses observed data which does not guarantee symmetry in the two inconsistent decisions. This lack of symmetry may be an advantage of the W-DC method for distributions that are not normal, however, it may be potentially increase error in the DC estimate because it will reflect anomalies in the frequency distribution of the observed data that may not be present if another random sample of examinees from the same population.

Method

This study used data sets from two simulated and three different operational exams to compare the similarity of the DC estimates produced by the LL-DC, PS-DC, and W-DC methods. Table 6 lists the characteristics of the five datasets, which includes an education, professional credentialing (PC), and healthcare exam. Dataset Sim1 was simulated to have a normal distribution of scores while Dataset Sim 2 was simulated to have a negatively skewed distribution. The operational datasets have some skewness, but Dataset B is the most skewed. In addition, Dataset A does not have a preset cut score since it is an admissions exam in which individual schools set their own standards. Datasets B and C are pass/fail exams. The reliabilities of the exam scores for the simulated datasets are 0.898 for the normally distributed scores and 0.841 for the skewed data distribution. The reliabilities of the scores for Datasets A-C range from 0.786 to 0.944. All datasets have sample sizes of at least 557 examinees.

For each dataset, the LL-DC, PS-DC, and W-DC methods are used to estimate the DC for every possible integer cut score. The results are compared to each other to observe similarities and differences in the results across the score distributions, in particular between LL-DC and the other two methods, and to make practical recommendations.

Results

Dataset Sim 1 – Normal Distribution

Figure 2 displays the results from the dataset simulated to have the exam scores normally distributed. Table 7 provides an excerpt from the results. Overall, all three methods produce DC estimates similar to each other across the score distribution. As seen in Table 7, the greatest difference in the DC estimates occurs when the cut score is set at 95 or 96. At these scores, the maximum difference in the DC estimate between the three methods is 0.020. Near the peak of this distribution, the PS-DC method has a slightly higher DC estimate compared to the LL-DC and W-DC methods. However, the lowest DC

Table 6. Description of Datasets

Dataset	Sim 1	Sim 2	A	B	C
Domain	Normal	Skewed	Education	PC	Healthcare
Exam Purpose	N/A	N/A	Admissions	License	License
Use of a pass/fail cut score	N/A	N/A	No	Yes	Yes
N examinees	1,000	1,000	6,785	1,509	557
N Scored Items	135	31	50	80	350
Mean	99.75	27.81	26.01	56.73	277.45
SD	9.83	2.89	10.41	7.86	29.12
Median	100	29	25	57	280
Mode	100	30	20	59	275
Excess Kurtosis	-0.02	-0.19	-0.66	1.55	-0.14
Skewness	-0.03	-0.93	0.35	-0.85	-0.46
Reliability	0.898	0.841	0.925	0.786	0.944
St. Error of Measurement	3.11	0.92	2.85	3.64	6.88

estimate for all three methods occurs at the mode of the dataset, i.e., a cut score of 100. This is an expected result since the DC values tend to be the lowest at the peak of a score distribution because this is the location of the highest chance of an inconsistent decision. Despite these small differences, all methods produce very similar DC estimates.

Dataset Sim 2 – Skewed Distribution

Figure 3 displays the results from the dataset simulated to have the exam scores negatively distributed. Table 8 provides an excerpt from the results. As seen in Table 8, the greatest difference in the three DC estimates across all possible scores value is 0.101. This occurs when the cut score is set at 30. The next greatest difference occurs at a cut score of 27. Here, the maximum difference occurs between the PS-DC and W-DC estimates and the magnitude of the DC difference was 0.069. For all other cut scores, the three methods produce DC estimates that are within 0.045 of each other. When comparing just W-DC and LL-DC, the greatest difference occurs at a cut score of 30 (difference = 0.082). All other differences are within 0.029 of each other. When comparing just PS-DC and LL-DC, the greatest difference occurs at a cut score of 27 (difference = 0.440). All other differences are within 0.037 of each other.

In general, one would expect the lowest DC estimate when the cut score is set near peak of the distribution, For Dataset Sim 2, the peak of the score distribution is at a score of 30 and only the W-DC method has the lowest DC estimate of all cut scores at this value. The lowest DC estimate for the LL-DC method occurs at the neighboring score of 29 and the lowest estimated by the PS-DC method occurs at a cut score of 28. These results suggest that the magnitude of the DC estimates by the LL-DC and PS-DC methods may be affected by the skewness of the dataset; however, the differences are small. In addition, the PS-DC estimates appear to be more affected by the skewness than the LL-DC method. As shown in Figure 3, the LL-DC and W-DC methods may be the better estimate with skewed because both of these methods attempt to adjust for this non-normality and these two methods do not include a normality assumption as does the PS-DC method.

Figure 2. DC estimates for each Possible Integer Cut Score and Frequency Distribution for Sim 1

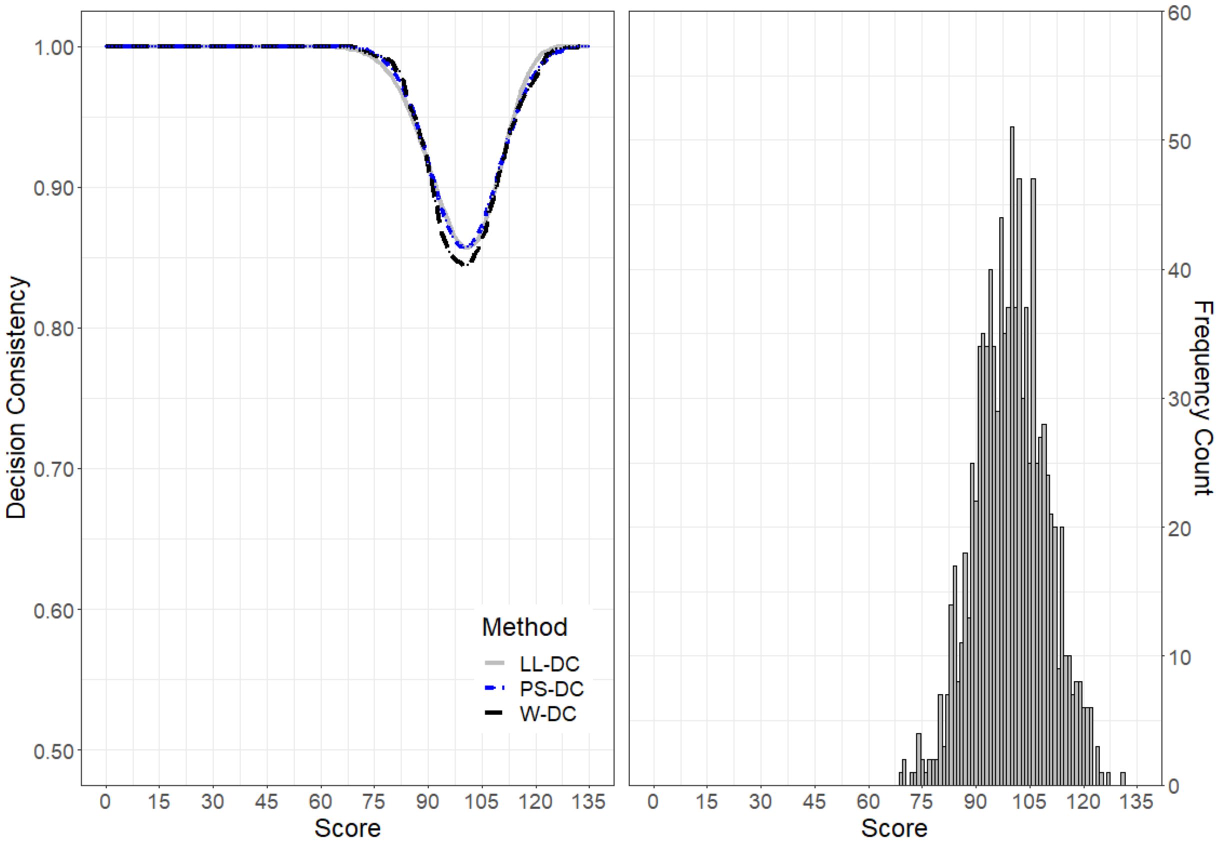
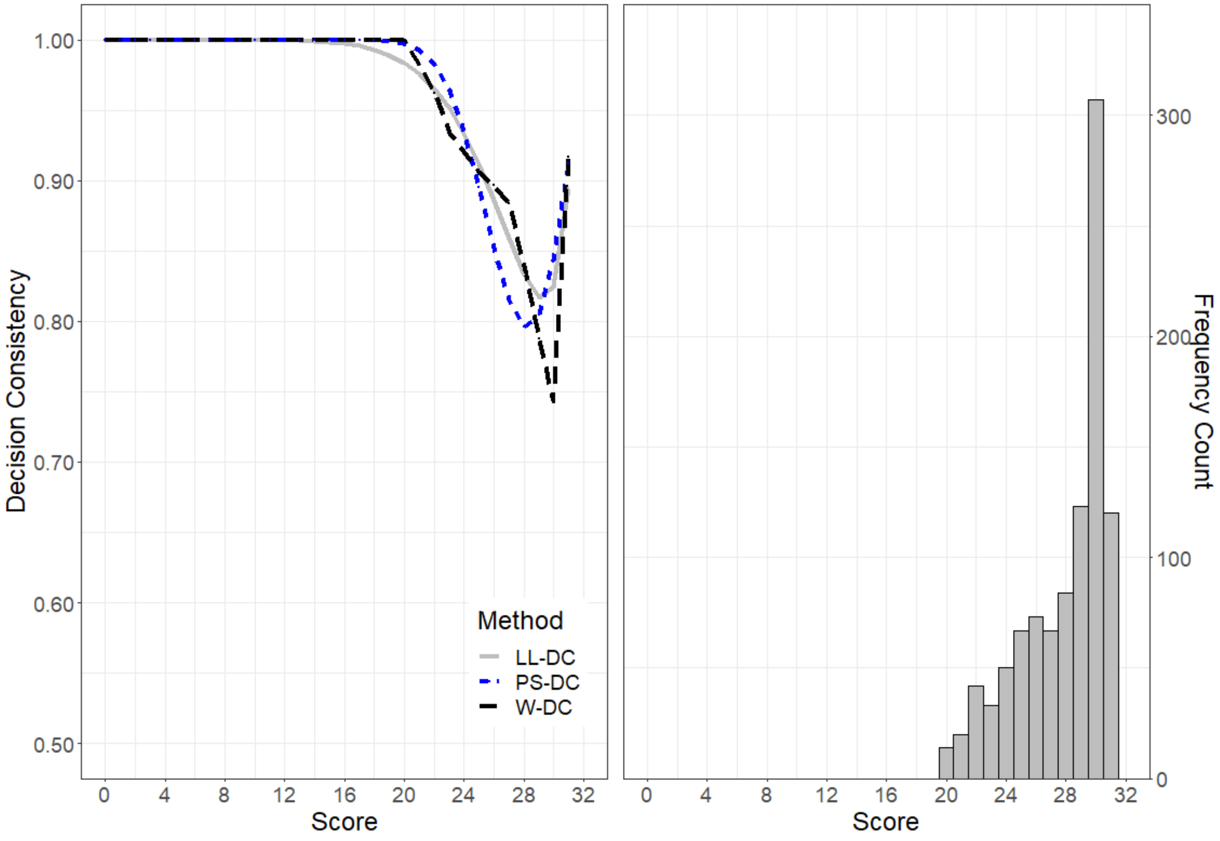


Table 7. Excerpt of the DC estimates Across the Score Distribution for Dataset Sim 1

Scores	N	LL-DC	PS-DC	W-DC	Max. Difference
0-91	197	0.909 to 1	0.909 to 1	0.905 to 1	0.013
92	35	0.901	0.900	0.890	0.011
93	34	0.892	0.892	0.877	0.016
94	40	0.886	0.884	0.867	0.019
95	34	0.879	0.876	0.859	0.020
96	29	0.873	0.870	0.853	0.020
97	44	0.867	0.865	0.851	0.016
98	35	0.862	0.861	0.847	0.016
99	37	0.860	0.858	0.847	0.013
100	51	0.857	0.857	0.844	0.013
101	37	0.857	0.857	0.844	0.013
102-135	427	0.858-1.000	0.859-1.000	0.847-1.000	0.012

*Shaded rows indicate the scores with the greatest value in the "Max. Difference" column.

Figure 3. DC estimates for each Possible Integer Cut Score and Frequency Distribution for Sim 2



Dataset A – Education

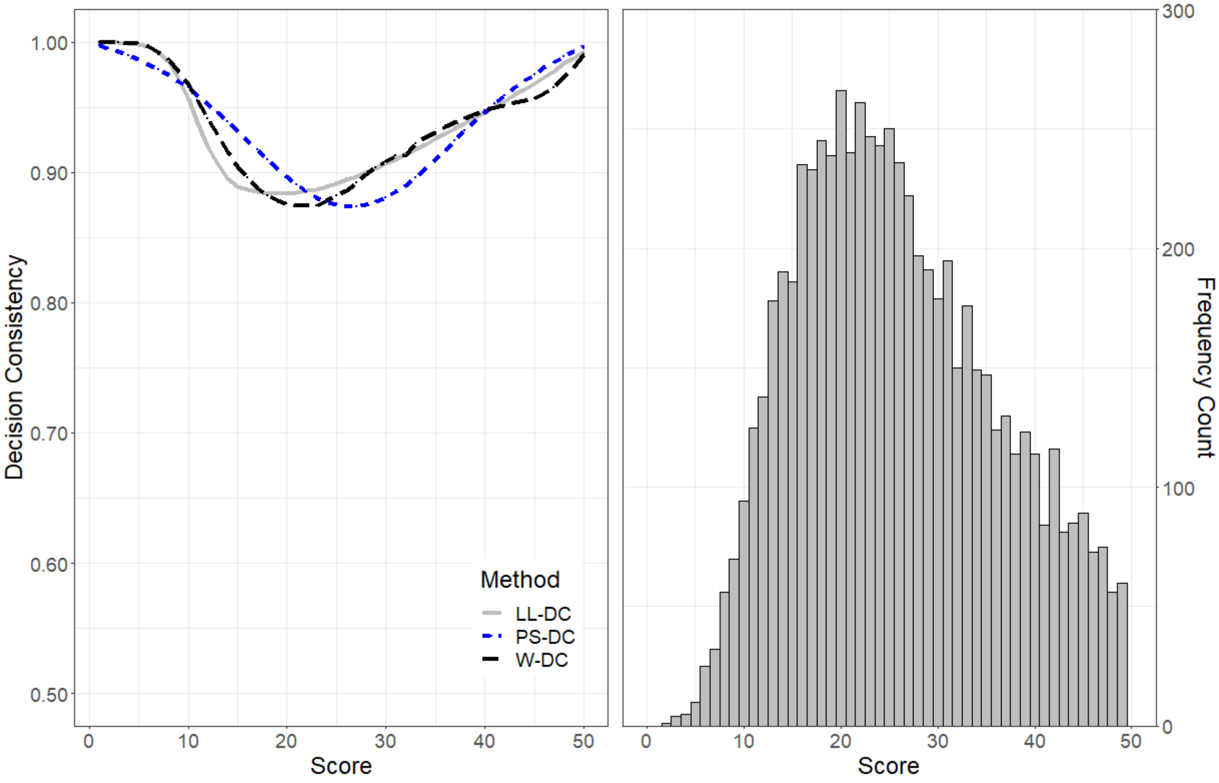
Figure 4 displays the results from Dataset A, which has a slight positive skewness and slight negative kurtosis. Table 9 provides an excerpt from the results. As seen in Table 9, the greatest difference in the three DC estimates across all possible scores value is 0.044. This occurs when the cut score is set at 14. For all cut scores set at 18 or higher (where most cut scores would likely be set), the three methods produce DC estimates that are within 0.028 of each other.

Table 8. Excerpt of the DC estimates Across the Score Distribution for Dataset Sim 2

Scores	N	LL-DC	PS-DC	W-DC	Max. Difference
0-25	226	0.912-1.000	0.896-1.000	0.906-1.000	0.031
26	73	0.886	0.852	0.897	0.045
27	67	0.859	0.815	0.884	0.069
28	84	0.833	0.796	0.839	0.043
29	123	0.817	0.804	0.788	0.029
30	307	0.825	0.844	0.743	0.101
31	120	0.894	0.912	0.916	0.022

*Shaded row indicates the score with the greatest value in the "Max. Difference" column.

Figure 4. DC Estimates for each Possible Integer Cut Score and Frequency Distribution for Dataset A



The W-DC and LL-DCs method produce the lowest DC estimates near the mode of the score distribution, i.e., at a score of 22 for W-DC and a score of 19 for LL-DC. The PS-DC method has the lowest DC estimate at a score of 26, which is near the mean of the dataset. These results all seem reasonable because a high density of scores at or near the cut score increases the number of examinees that may have an inconsistent decision. On the other hand, if the cut score are in a sparser area of the score distribution, then the probability of an inconsistent decision decreases and the DC estimate is expected to be higher.

Overall, the W-DC and LL-DC methods produce DC estimates similar to each other across the score distribution. The PS-DC method tends to produce slightly higher DC estimates when the cut score is set at values between 11 and 21 as well as between 41 and 50. As shown in Figure 4, the skewness of the dataset suggests that either the W-DC or the LL-DC method may be the better estimate in this situation because their methods both attempt to adjust for this non-normality.

Dataset B – Licensure Exam

Figure 5 displays the results from Dataset B, which has a negatively skewed and positively kurtotic score distribution. Table 10 provides an excerpt from the results. As seen in this table, the greatest difference in the DC estimates across the three methods is 0.063. This occurs when the cut score is set at 63, which is near the mode of the dataset. At this score point, the LL-DC, PS-DC, and W-DC methods estimate the DC to be 0.78, 0.84, and 0.80, respectively. Cut scores near this peak in the distribution (e.g., 58 through 65) also have greater differences in the DC estimates across the methods compared to those scores earned with less frequency. The differences in the DC estimates of cut scores ranging from 58 to 65 is greater than 0.04, with the PS-DC method consistently producing the highest DC estimates across the three methods. The W-DC method has the lowest DC estimate from cut scores of 58 to 61 and LL-DC has the lowest DC estimate from cut scores of 62 to 65. The differences in the DC estimates for cut scores set at less than 58 are within approximately 0.038 of each other and those above 65 are within 0.032 of each other. As with Dataset A, the PS-DC method is influenced more by the skewness in the dataset than the other methods. However, the three DC methods produced similar DC estimates for most of the possible cut scores.

Dataset C – Healthcare

Figure 6 displays the results from Dataset C. In this dataset, the score distribution is slightly negatively skewed with kurtosis similar to that of a normal distribution. Table 11 provides an excerpt from the results. As seen in this table, the greatest difference in the three DC estimates across all possible score values is 0.025. This value occurs when the cut score is set at 307, 308, 316, 317, or 318. At all five of these scores, the LL-DC method produces the lowest DC estimate and the PS-DC produces the highest. However, the methods all produce relatively similar DC estimates.

With Dataset C having a distribution close to that of a normal distribution, it is not surprising that the three methods show similar DC estimates across all possible cut scores. However, this dataset illustrates

Table 9. Excerpt of the DC Estimates Across the Score Distribution for Dataset A

Scores	n	LL-DC	PS-DC	W-DC	Max. Difference
1-9	0-203 per score	0.972-1.000	0.970-1.000	0.978-1.000	0.014
10	94	0.956	0.965	0.968	0.011
11	125	0.938	0.959	0.955	0.021
12	138	0.920	0.953	0.942	0.033
13	178	0.906	0.946	0.929	0.040
14	190	0.895	0.939	0.916	0.044
15	186	0.889	0.932	0.905	0.043
16	235	0.887	0.925	0.896	0.038
17	233	0.885	0.917	0.888	0.032
18	245	0.884	0.910	0.882	0.028
19	239	0.883	0.903	0.879	0.024
20	266	0.884	0.896	0.876	0.020
21	240	0.885	0.890	0.875	0.016
22	261	0.886	0.885	0.874	0.012
23	247	0.887	0.881	0.874	0.013
24	243	0.889	0.877	0.878	0.012
25	250	0.891	0.875	0.882	0.016
26	236	0.894	0.874	0.886	0.021
27	222	0.897	0.874	0.892	0.023
28	197	0.899	0.875	0.898	0.025
29	191	0.903	0.877	0.903	0.026
30	179	0.907	0.881	0.908	0.028
31-50	46-195 per score	0.910-0.993	0.885-0.997	0.913-0.990	0.028

*Shaded row indicates the score with the greatest value in the “Max. Difference” column.

Figure 5. DC Estimates for each Possible Integer Cut Score and Frequency Distribution for Dataset B (Licensure Exam)

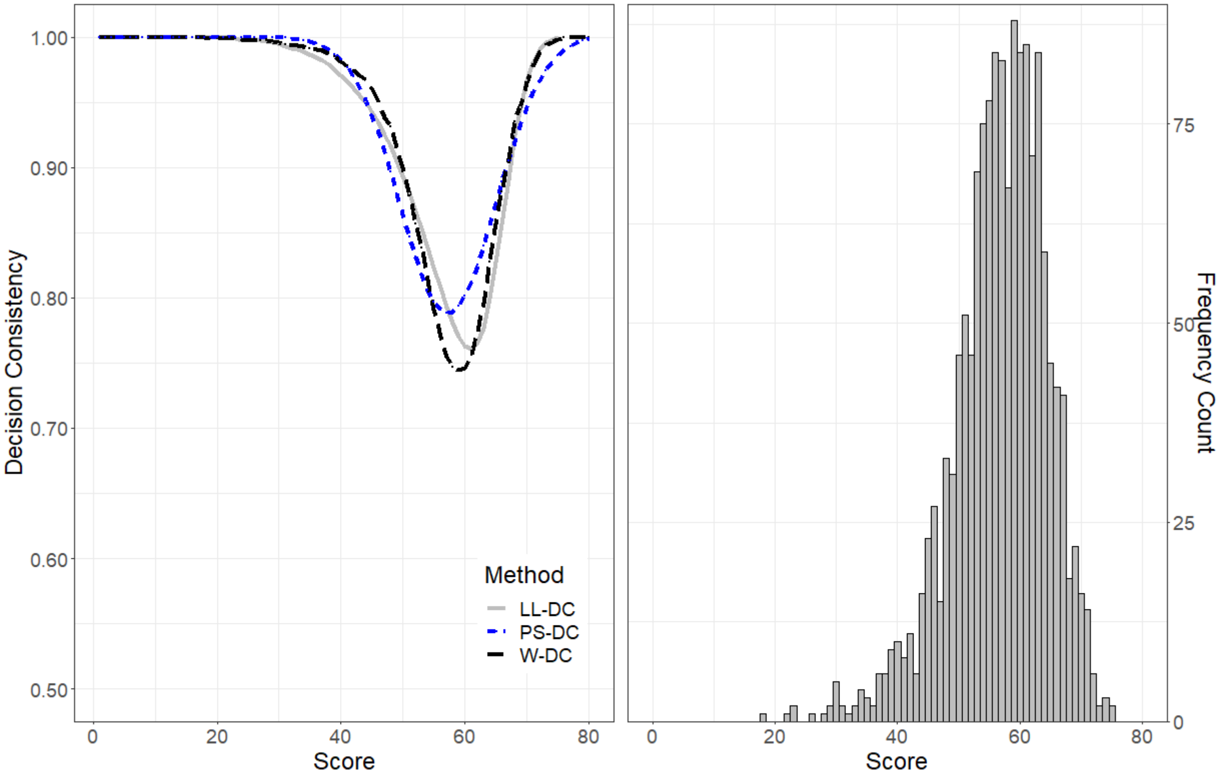
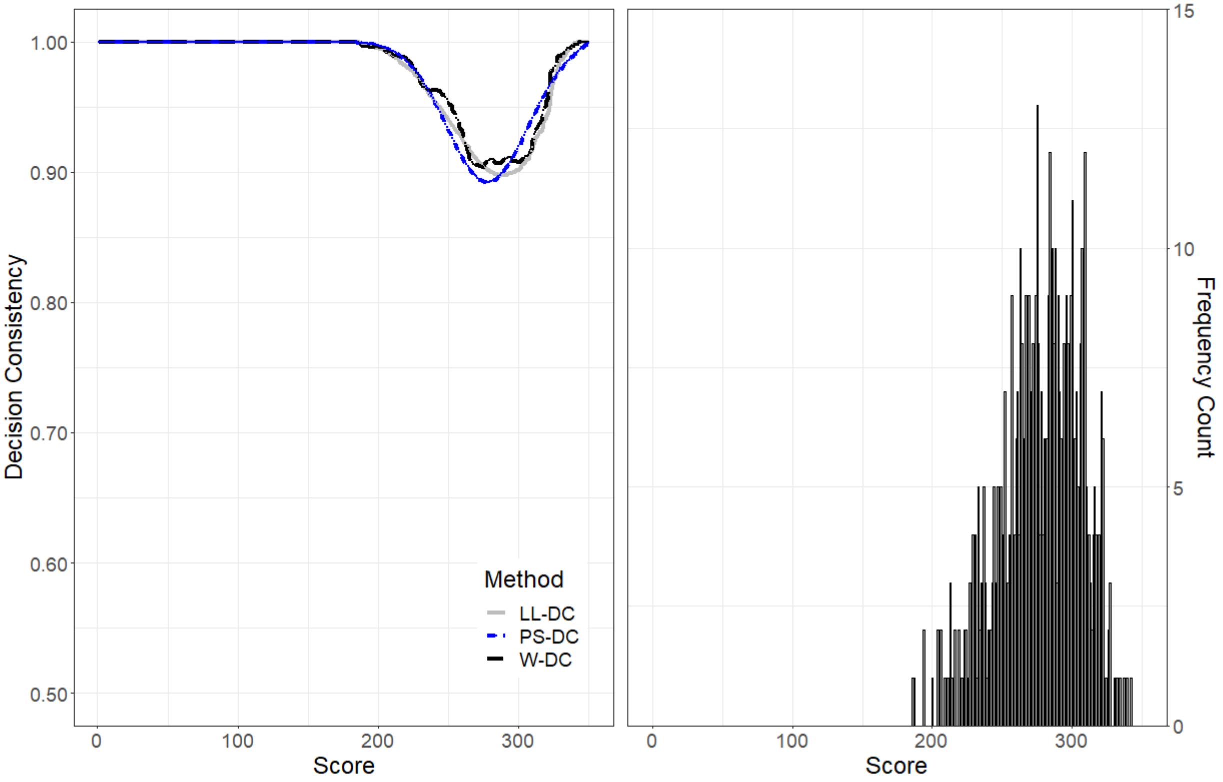


Table 10. *Excerpt of the DC Estimates Across the Score Distribution for Dataset B*

Scores	n	LL-DC	PS-DC	W-DC	Max. Difference
1-55	0-78 per score	0.823-1.000	0.797-1.000	0.791-1.000	0.037
56	84	0.808	0.791	0.770	0.038
57	83	0.793	0.788	0.755	0.038
58	67	0.780	0.789	0.747	0.042
59	88	0.770	0.793	0.745	0.049
60	84	0.763	0.801	0.746	0.055
61	85	0.761	0.812	0.756	0.056
62	71	0.766	0.825	0.775	0.059
63	84	0.777	0.840	0.796	0.063
64	59	0.800	0.856	0.828	0.056
65	45	0.827	0.872	0.856	0.045
66	42	0.857	0.889	0.881	0.032
67	41	0.890	0.905	0.907	0.016
68	18	0.925	0.920	0.935	0.015
69	22	0.946	0.934	0.948	0.014
70-80	0-16 per score	0.966-1.000	0.946-0.999	0.965-1.000	0.024

*Shaded row indicates the score with the greatest value in the “Max. Difference” column.

Figure 6. DC Estimates for each Possible Integer Cut Score and Frequency Distribution for Dataset C (Licensure Exam - Healthcare)



the effect of using only the observed data to estimate the DC, as is done in the W-DC method. Seen in Figure 6, the wiggly DC estimates at the high frequency score range is not ideal and depends on the observed score distribution. In these cases, either a smoothed version of the W-DC estimate or an estimate based on theoretically distributed data (such as the LL-DC or PS-DC methods) may be a better DC estimate.

Overall Comparison

One of the purposes of this study was to compare how well the computationally and conceptually simpler PS-DC and W-DC estimates compared to the LL-DC estimates. For the dataset simulated to have a normal distribution of total scores, the W-DC estimates are within 0.020 of the LL-DC estimates and the PS-DC estimates are within 0.009. For the dataset simulated to have a skewed distribution of total scores, the W-DC estimates are within 0.082 of the LL-DC estimates and the PS-DC estimates are within 0.044. For Dataset A, B, and C, the W-DC estimates are within 0.023, 0.039, and 0.019 of the LL-DC estimates, respectively, and the PS-DC estimates are within 0.044, 0.063, and 0.025, respectively.

The greatest differences observed in the above results tended to occur when the cut score was set at the score with the highest frequency. While the differences are more noticeable at these score points, these differences from the LL-DC estimate is simply an indication that the given method deviates from the estimates calculated by the LL-DC method. It is possible that the DC estimates computed via the PS-DC or the W-DC method be more accurate at these score points. In particular, the W-DC method produced the expected results of having the lowest DC across the score distribution at the peak of the distribution. The PS-DC and LL-DC methods had the lowest DC near the peak of the distribution, but not necessarily at the score associated with the peak.

Discussion

This study compared the DC estimates for all possible cut scores of two simulated exams and three operational exams using three different DC methods: LL-DC, PS-DC, and W-DC. Overall, the DC estimates across all datasets ranged from 0.745 to 1.000. A DC estimate of 0.745 practically means that if examinees

Table 11. Excerpt of the DC Estimates Across the Score Distribution for Dataset C

Scores	n	LL-DC	PS-DC	W-DC	Max. Difference
1-306	0-13 per score	0.898-1.000	0.892-1.000	0.904-1.000	0.023
307	10	0.910	0.935	0.913	0.025
308	9	0.912	0.937	0.915	0.025
309	12	0.916	0.939	0.917	0.023
310	5	0.919	0.942	0.925	0.023
311	2	0.924	0.944	0.929	0.020
312	4	0.926	0.946	0.931	0.020
313	3	0.927	0.948	0.934	0.021
314	2	0.929	0.950	0.936	0.022
315	4	0.931	0.952	0.937	0.024
316	5	0.931	0.955	0.940	0.025
317	4	0.934	0.957	0.944	0.025
318	3	0.937	0.959	0.948	0.025
319	4	0.941	0.961	0.950	0.024
320	3	0.943	0.962	0.955	0.024
321	7	0.947	0.964	0.958	0.024
322	6	0.951	0.966	0.967	0.022
323	1	0.959	0.968	0.977	0.018
324	1	0.967	0.970	0.979	0.011
325	1	0.970	0.971	0.980	0.010
326	2	0.972	0.973	0.981	0.009
327	3	0.975	0.974	0.984	0.010
328-350	0-1 per score	0.979-1.000	0.976-1.000	0.989-1.000	0.013

*Shaded rows indicate the scores with the greatest value in the "Max. Difference" column.

were to retake a parallel form of the exam with no memory of their first attempt and under identical conditions to their first attempt, then 25.5% would *not* have the same pass/fail decision as in their first attempt.

One of the purposes of this study was to determine if the less complex methods of PS-DC and W-DC produced results similar to the more widely used and accepted, yet more complex, LL-DC method. In comparing the W-DC and PS-DC methods to LL-DC, the PS-DC estimates were closer to the LL-DC estimates for the simulated datasets, but the W-DC estimates were closer to the LL-DC estimates for the operational datasets. In particular, the PS-DC estimates were closer to the LL-DC estimates for 65-71% of the scores in the simulated datasets and the W-DC estimates were closer to the LL-DC estimates for 61-74% of the scores in the operational datasets.

The greatest difference in the DC estimates of the simulated datasets was 0.101 and occurred in the skewed dataset at the mode of the distribution. At this score of 30 out of 31, the LL-DC method estimated the DC index to be 0.825, while the PS-DC and W-DC methods estimated the index to be 0.844 and 0.743, respectively. All other DC estimates for the skewed datasets were within 0.029 of each other. This result highlights the effect of datasets with skewed score distributions as well as differences in the DC estimate based on the location of the cut score. Theoretically, the lowest DC index would occur at the mode of the distribution. This happened when the W-DC method was applied to this dataset, but not during the application of the PS-DC and LL-DC methods.

In the operational datasets, the greatest difference in the DC estimates occurred in Dataset B, which was also the most skewed operational dataset. This difference occurred at a score with high frequency, i.e., score of 63 out of 80. At this score, the LL-DC method estimated the DC index to be 0.777 while the PS-DC and W-DC methods estimated the index to be 0.839 and 0.796, respectively. Due to the assumptions of normality in the PS-DC method, it is not unexpected that the PS-DC tended to be less aligned with the other two methods at the scores with highest frequency. However, the PS-DC method seemed reasonably well aligned with the W-DC and LL-DC estimates at the scores earned with less frequency.

The three DC methods produced similar DC estimates across the score distribution for the normally (or close to normally) distributed simulated dataset and two operational datasets, i.e. all had values within 0.044 of each other. The W-DC method showed small fluctuations in the DC estimates across the score distributions when the cut score was set at scores earned with higher frequency. These fluctuations mirror those found in the observed data. The PS-DC and LL-DC had smoother distributions of the DC estimate for cut scores set across the score distribution. A smoothing adjustment could be applied to the W-DC method. This may make it more aligned with the other methods, but may not necessarily make it a more accurate estimate. More research would be needed to determine the effects of applying a smoothing technique to this calculation. If such research indicated an improvement in the W-DC estimates, then the tradeoff for adding complexity to the model in exchange for a magnitude of improved accuracy would need to be evaluated.

Table 12 summarizes recommendations based on the results from this study. As evidenced by both the simulated and operational datasets and of the three methods compared, the W-DC and LL-DC methods are recommended for skewed data. The PS-DC method may be used for skewed data but will likely overestimate the DC value if the cut score is set at a score frequently earned by examinees. All methods are recommended for normally distributed data; however, the W-DC method is sensitive to small deviations in the frequency distribution of scores in the observed data.

In deciding which method to apply, a user should consider their own ability as well the intended audience. If the goal is to for a measurement professional to report the DC estimate based on the scores on an exam without having to explain how the method works to a client or other non-measurement professional, then the LL-DC method is a widely used and accepted method that would suffice for this purpose. Most measurement professionals have the background knowledge to compute this value with an available computer program. The W-DC method could also be used, but it is a new method that has not yet been widely applied in the field. The PS-DC method could also be used, but this method would be better suited for data following a normal distribution.

If the goal is for a measurement professional to report the DC estimate and explain in layman terms how the method works to a client or other non-measurement professional, then a user may opt for the simpler W-DC or PS-DC methods. The user could check the similarity of these values against the LL-DC estimate, if desired. However, a comparison would only inform the user if the values are similar and not which value is more accurate. Again, the PS-DC method would not be recommended for skewed datasets.

Finally, if the goal is for a non-measurement professional to compute the DC estimate for a set of exam scores, then the W-DC method is recommended. This method is simple to compute, does not require a computer program, and is less complex conceptually. If the user has some measurement or statistical background and the data is normally distributed, the PS-DC method is also an option.

Overall, the results of this study show that the three DC methods provide similar DC estimates for normally distributed datasets. The W-DC and LL-DC methods produce similar results for skewed datasets. While the results of this study did not evaluate the accuracy of the methods, it is important to remember that these values are *estimates* of the decision consistency. The similarity of the methods suggest that they produce reasonable estimates. If desired, a user could compute and report multiple DC estimates (including methods using IRT) to have more confidence in the value.

Guidance on an Acceptable DC Index

The industry provides little guidance on what constitutes an acceptable value for decision

Table 12. Recommended Method for Estimating DC from a Representative Sample of Data

★ = Recommended; ✓ = Recommended with reservations; × = Not recommended

	PS-DC	**LL-DC**	**W-DC**
Skewed	×	★	★
Normal	★	★	✓[1]

[1]May have small dips and peaks in the DC estimates to reflect any dips and peaks in the observed score distribution

This, in part, is due to the value being influenced by the reliability of the exam scores on the form as well as the frequency distribution of scores near the cut score. In particular, if there is a high frequency of examinees scoring at or near the cut score, then there is a greater chance of error in the pass/fail decision. Therefore, the DC estimate would be lower at the cut score than if the cut score were located at a score in which fewer examinees scored. However, since credentialing programs are often trying to write items to separate those that are and are not minimally qualified, there are often many items written to target the ability of a minimally qualified examinee. This is also the target ability of the cut score. This, in turn, creates a higher frequency of examinees scoring at the cut score and causes a decrease in the magnitude of the DC index. This result does not imply a poor exam. However, increasing the reliability of the exam scores will help the magnitude of the DC estimate be as high as possible.

Of the little guidance that is offered on an acceptable value for the DC estimate, Subkoviak (1988) provided the following general rule: "Tests used to make serious decisions should be sufficiently long to guarantee an agreement coefficient exceeding 0.85. Higher values can be expected as the relative proportions of masters and nonmasters become more dissimilar. . . A full-period classroom test should guarantee an agreement coefficient of at least 0.75" (p. 52). In general, the authors agree with this guidance, but believe it is not always realistic in professional credentialing exams. The determination of an acceptable DC estimate should be a programmatic policy decision that answers the question: What is an acceptable percentage of agreement of observed pass/fail decisions with pass/fail decision that would be obtained from going back in time and having the examinee take the exam again without memory of taking it the first time? The answer to this question should be informed by data related to the exam program, sample size, the reliability of the exam scores, and the frequency distribution of the data.

Many exam programs will likely desire to select one DC method and apply it to all situations (much like how Cronbach's alpha is commonly applied). This study illustrates that each DC method has its strengths, weaknesses, and limitations, and that one method does not work for all situations. However, the "true" DC is an unknown quantity and this study shows that the LL-DC, PS-DC, and W-DC methods produce very similar estimates most of the time. While the authors agree with Subkoviak's guidelines suggested above, the authors believe that users of the DC estimate should acknowledge and emphasize that the DC value is an *estimate* of the "true" DC value and is influenced by multiple factors, including the sample size, reliability of the exam scores, location of the cut score, and the score frequency distribution. Thus, while a DC estimate of at least 0.85 is a goal, there could be an empirical reason for the value to be lower. For example, if the peak of the score distribution is at the cut score, then the DC value may be lower than desired. This is potentially a justifiable result if the peak is due to a large number of items on the exam focused at the ability level of the minimally qualified candidate. If the cut score were set in a sparser area of the score distribution, then a higher DC may be expected. The DC is also affected by the reliability of the exam scores. Lower (or higher) reliability values will lead to lower (or higher) DC values. Since reliability is often a function of the distribution of examinees, very homogeneous populations will have exam scores with lower reliabilities, and therefore likely lower DC, than very heterogeneous populations. So while low reliabilities are not desired, it may explain a lower DC estimate. Related to reliability and as suggested by Subkoviak (1988), the DC guidelines assume that an exam has a sufficiently large number of items. The above guidelines may need to be adjusted for shorter exams and for the relative heterogeneity of the test taking population.

Limitations and Opportunities for Further Research

This study compared three different methods for estimating DC using two simulated datasets and three operational datasets. Additional simulation studies comparing the different methods would be beneficial and may provide further insight into the effects of sample size, skewness, and kurtosis on the different methods. In addition, simulation studies comparing these methods to "true" DC indices would help determine in what situations one method is more accurate than another.

This study also limited itself to three classical test theory methods for estimating DC: LL-DC, PS-DC, and W-DC. There are other methods for estimating DC, including item response theory methods, that could be compared. Such comparisons (e.g., Stoeger & Skorupski, 2023) allow for more recommendations on when to use or not use a certain method, the comparability of different methods, and may also provide additional guidance on acceptable DC values given a score distribution, sample size, location of cut score, and the reliability of the exam scores.

Conclusion

This study compared LL-DC, PS-DC, and W-DC methods for estimating DC. The results indicated that the W-DC and LL-DC methods produce the most reasonable results for skewed data. While all methods produce reasonable results for normally distributed data, the W-DC method noticeably reflects the peaks and valleys in the observed.

Overall, the DC method selected by a user depends on their own measurement knowledge and ability as well as the intended audience. On one extreme, if the complexity of the model is not of concern and it is not necessary to explain how the model works to a non-measurement audience, then the LL-DC method may be the preferred choice. On the other extreme, if a simpler method that is more straightforward to explain is desired, then the W-DC method may be the preferred choice.

The authors support the recommendation of having a DC estimate of at least 0.85 for high stakes exams, but also believe there are reasonable explanations as to why the DC estimate may be lower. If a DC estimate is lower than the recommended guideline, then it is recommended that one look closer at the dataset, the nature of the exam, and the nature of the testing population to determine reasons for the lower value.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. AERA: Washington, DC.

Alger, S. (2016). Is this reliable enough? Examining classification consistency and accuracy in a criterion-referenced test. International Journal of Assessment Tools in Education, 3, 137-150. https://doi.org/10.21449/ijate.245198

Brennan, R. L. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy (Version 1.0).* (CASMA Research Report No. 9). [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs

Breyer, F. J., & Lewis, C. (1994). Pass-fail reliability for tests with cut scores: A simplified method. ETS Research Report Series (Report No. RR-94-39). https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.1994.tb01612.x

Cicchetti D.V. & Feinstein A.R. (1990). High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology, 43(6), 551–558. https://doi.org/10.1016/0895-4356(90)90159-M

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20(1), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213–220. https://doi.org/10.1037/h0026256

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334.

Deng, N. (2011). Evaluating IRT- and CTT-based methods of estimating classification consistency and accuracy indices from single administrations. Open Access Dissertations. 452. https://scholarworks.umass.edu/open_access_dissertations/452

Feinstein A.R., & Cicchetti D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Gwet, K. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment, 1*. Available online: https://www.agreestat.com/papers/kappa_statistic_is_not_satisfactory.pdf

Haakstad, H. E. (2022). betafunctions: Functions for Working with Two- And Four-Parameter Beta Probability Distributions. R package version 1.6.1.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*(4), 253-264.

Li, S. (2006). Evaluating the consistency and accuracy of proficiency classifications using item response theory. https://search.proquest.com/openview/b003729f52536da71818a25195dd709b/1?pq-origsite=gscholar&cbl=18750&diss=y

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179-197.

Marshall, J. L., & Haertel, E. H. (1975). A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement. (ED118618). ERIC. https://files.eric.ed.gov/fulltext/ED118618.pdf

McDonald R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum.

Mellenbergh, G. J., & van der Linden, W. (1979). The internal and external optimality of decisions based on tests. Applied Psychology Measurement, 3(2), 257-273.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17*(4), 359-368.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at http://www.R-project.org/

Stoeger, J. N., Skorupski, W. (2023). A comparison of methods to evaluate the consistency of cutscore decisions. Presented at the National Council on Measurement in Education conference. Chicago, IL.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*(4), 265-276,

Wan, L., Brennan, R. L., & Lee, W. (2007). *Estimating classification consistency for complex assessments.* (CASMA Research Report No. 22). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. https://education.uiowa.edu/sites/education.uiowa.edu/files/documents/centers/casma/publications/casma-research-report-22.pdf

Wolkowitz, A. A. (2021). *A Computationally Simple Estimate for Decision Consistency. Journal of Educational Measurement, 58*(2), 388-412.

Young, M. J., & Yoon, B. (1998). Estimating the consistency and accuracy of classifications in a standards-referenced assessment (ED426067). ERIC. https://files.eric.ed.gov/fulltext/ED426067.pdf

Citation:
Wolkowitz, A. A., & Smith, R. (2024). A practical comparison of decision consistency estimates. *Practical Assessment, Research, & Evaluation*, 29(6). Available online: https://doi.org/10.7275/pare.2023

Corresponding Author:
Amanda A. Wolkowitz
Data Recognition Corporation

Email: awolkowitz [@] datarecognitioncorp.com

Appendix A.

Code for implementing W-DC in R (R Core Team, 2022)

```
#*******************************************************************************
#Purpose: This program computes W-DC
#Required Inputs:
#   1. Total score file (with header row)
#   2. Reliability of exam.
#   3. Maximum possible score on exam.
#   4. Cut score
#Output: W-DC for given cut score
#Last updated: 01/24/2022
#*******************************************************************************

#*************Functions*******************
#Define function for first half of DC  - Values less than cutscore
below_cut_function <- function(score) {

  # If score < cut score AND  within lower bound (LB) and upper bound (UB) then count number of failing examinees with scores between score - 95%CI (inclusive) and cut score (exclusive)
  a_LBindex <- max(which(mydata_df$Score == 0), which(mydata_df$Score == score - plusminusCI))
  a_UBindex <- min(which(mydata_df$Score == score + plusminusCI), which(mydata_df$Score == cutscore - 1))
  a <- sum(mydata_df$Freq[a_LBindex:a_UBindex])

  #Count total number of examinees scoring between score +/- SEM (inclusive)
  CI_UBindex <- min(which(mydata_df$Score == score + plusminusCI), which(mydata_df$Score == maxpossible))
  a_total <- sum(mydata_df$Freq[a_LBindex:CI_UBindex])

  # Divide a by a_total
  if(a_total == 0) {
    pFail <- 1
  } else {
    pFail <- a/a_total
  }

  # Multiply by number of examinees scoring that score to determine number of examinees with  consistent fail
  c <- mydata_df$Freq[which(mydata_df$Score == score)]
  nFail <- pFail * c

  #Add number of consistent fail decisions to score table
  DC[score+1] <-nFail
}

#Define function for second half of DC  - Values greater than cutscore
above_cut_function <- function(score) {

  # If score >= cut score AND  within LB and UB then Count number of passing examinees with scores between cut score and score + SEM (inclusive)
  b_LBindex <- max(which(mydata_df$Score == score - plusminusCI), which(mydata_df$Score == cutscore))
  b_UBindex <- min(which(mydata_df$Score == maxpossible), which(mydata_df$Score == score + plusminusCI))
  b <- sum(mydata_df$Freq[b_LBindex:b_UBindex])

  #Count total number of examinees scoring between score +/- SEM (inclusive)
  CI_LBindex <- max(0, which(mydata_df$Score == score - plusminusCI ))
  b_total <- sum(mydata_df$Freq[CI_LBindex:b_UBindex])

  # Divide b by b_total
  if(b_total == 0) {
    pPass <- 1
  } else {
    pPass <- b/b_total
  }

  # Multiply by number of examinees scoring that score to determine number of examinees with  consistent pass
  c <- mydata_df$Freq[which(mydata_df$Score == score)]
  nPass <- pPass * c

  #Add probability of consistent decision to score table
  DC[score+1] <- nPass

}
#************END FUNCTIONS******************

#Open data
library(readxl)
mydata <- read_excel("LOCATION OF DATA", sheet = "SHEET NAME")

#order total scores in decreasing order
mydata <-mydata$COLUMN_HEADER

#Get input from user for reliability, max possible score, and cut score
myrel <- readline(prompt="Enter reliability: ")
maxpossible <- as.numeric(readline(prompt="Enter maximum possible score: "))
cutscore <- as.numeric(readline(prompt="Enter cut score as positive integer: "))

#Create table for DC values
WDCtable <-vector(mode = "double", length = maxpossible)

#Error message if cut score out of bounds of possible scores
if(cutscore <= 0) {
  stop("Entered cut score is less than or equal to zero. This program requires a cut score between 0 and the maximum possible score.")
} else if (cutscore > maxpossible) {
  stop("You entered a cut score greater than the maximum possible score.")
}

#Convert user input to numeric
myrel <- as.numeric(myrel)
maxpossible <- as.numeric(maxpossible)
cutscore <- as.numeric(cutscore)

#Compute SEM
SEM <- sd(mydata, na.rm = TRUE)*sqrt(1-myrel)

#Compute +\-95% CI of SEM and round result up
plusminusCI <- SEM * qnorm(0.975, mean = 0, sd = 1)
plusminusCI <- ceiling(plusminusCI)

#Lower and upper bounds for CI. Max needed in LB to avoid negative LB and min needed in UB to avoid UB being greater than max
possible score
LB <- max(cutscore - plusminusCI, 0)
UB <- min(cutscore + plusminusCI, maxpossible)

#Create frequency table from 0 to possible scores.
mydata_table <- table(factor(mydata, levels = c(0:maxpossible)))
mydata_df <- as.data.frame(t(mydata_table)) #transpose
colnames(mydata_df)= c("A", "Score", "Freq")
```

```
as.numeric(mydata_df$Score)
as.numeric(mydata_df$Freq)

#Identify location in vector for LB and UB value
LBindex <- min(which(mydata_df$Score == LB))
UBindex <- max(which(mydata_df$Score == UB))

#Count frequency below and below 95%CI
BelowCI <- sum(mydata_df$Freq[0:(LBindex-1)]) #index 0 to one below LBindex
AboveCI <- sum(mydata_df$Freq[(UBindex+1):(maxpossible+1)]) #one above UBindex to index of max possible score

#Define DC vector
DC <- vector(mode = "double", length = length(mydata_df$Score))

#Apply DC functions
x <- seq(LB, cutscore - 1, 1)
y <- seq(cutscore, UB, 1)

DC[x+1] <- sapply(x, below_cut_function)
DC[y+1] <- sapply(y, above_cut_function)


#Above and below CI, DC = 1 so list frequency (100% of examinees will have consistent decision)
DC[1:(LBindex-1)] <- mydata_df$Freq[1:(LBindex-1)]
DC[(UBindex):length(mydata_df$Score)] <- mydata_df$Freq[(UBindex):length(mydata_df$Score)]

#Compute W-DC (sum of the number of examinees with a consistent decision divided by the total number of examinees)
WDC = sum(DC)/sum(mydata_df$Freq)

#Display result
cat("W-DC = ", WDC)
```