CHAPTER SIX

Investigating reliability in criterion-referenced tests

II. Conceptual exercises

A. Matching		B. True or false	
1.	c	1.	F
2.	g	2.	T
3.	d	3.	T
4.	f	4.	T
5.	i	5.	F
6.	a	6.	T
7.	h	7.	F
8.	b	8.	F

C. Brief answers

1. The most appropriate agreement index would be the estimated proportion of agreement, $\ddot{p}_o$. This agreement treats all misclassifications as equally serious, and can be used with multiple ratings. Coefficient kappa ($\ddot{\kappa}$) is likely not to be appropriate for several reasons: First, it may overcorrect for chance agreements; second, the teachers most likely want to know how much agreement there is for any reason, not just agreement that is due to the measurement procedure; and third, the teachers are not likely to want to generalize their estimate beyond their own classes.

2. Either the phi lambda, Φ_λ or the kappa squared ($\kappa^2{}_{(X,T_X)}$) would be appropriate because these squared-error-loss agreement indices treat misclassifications far from the cut score as more serious than those near the cut score. Phi lambda might be preferred in this situation, because it is marginally easier to calculate since it does not require an NR reliability estimate.

3. Since the DOE has set absolute cut scores for these two measures, the most appropriate agreement index would be the kappa coefficient ($\ddot{\kappa}$), which treats all misclassifications as equally serious. This would be preferred over the estimated

proportion of agreement ($\ddot{p}_o$) in this case, since the DOE will need to be able to generalize our estimate to all students in the state who will be assessed.

4. The phi would be the most appropriate, because you could estimate this for different lengths of tests to determine the length needed to achieve a dependability of .80.

5. Step 1: estimate the internal consistency reliability for the test.

Step 2: use Equation 6.1 in the Textbook to estimate phi.

Step 3: increase k in Equation 6.1 and recalculate phi.

Step 4: repeat step 3 until you obtain a phi coefficient of .80

III. Hand calculations with small data sets

A. Using the data below, for Test 1 (X_1), calculate the CR dependability and agreement indices below:

X_1 statistics:

n = 20	$\bar{X} = 43.6$	s = 3.152
k= 50	$\bar{X}_p = \dfrac{43.6}{50} = .872$	$s_p = \dfrac{3.152}{50} = .063$

KR21 = .447

1. Calculate Φ, using KR21 as an estimate of internal consistency.

$$\Phi = \frac{\frac{ns^2{}_p}{n-1}(KR-21)}{\frac{ns^2{}_p}{n-1}(KR-21)+\frac{\bar{X}_p(1-\bar{X}_p)-s^2{}_p}{k-1}}$$

$$\Phi = \frac{\frac{20*.00397}{20-1}(.447)}{\frac{20*.00397}{20-1}(.447)+\frac{.872(1-.872)-.00397}{50-1}}$$

$$\Phi = \frac{\frac{.0794}{19}(.447)}{\frac{.0794}{19}(.447) + \frac{.107646}{49}}$$

$$\Phi = \frac{.00186381}{.00186381 + .002196857} = .459$$

2. Calculate SEM_{abs} and .95 confidence intervals (CIs) for scores of 47, 43, 40, and 37, where 40 is the cut score.

$$\text{SEM}_{\text{abs}} = \sqrt{\frac{.872(1-.872)-.00378}{50-1}} = \sqrt{\frac{.111616-.00378}{49}} =$$

$$\sqrt{.0022007} = .0469$$

Remember that this is scaled to the correct proportion, so we need to rescale this to the raw score scale by multiplying it by the number of items, which is 50.

$\text{CI}_{.95}$, X, 47 = 47 ± 1.96*(.0469*50) = 42.4038-51.5962, rounds to 42-52

$\text{CI}_{.95}$, X, 43 = 43 ± 1.96*(.0469*50) = -38.4038-47.5962, rounds to 38-48

$\text{CI}_{.95}$, X, 40 = 40 - 1.65*(.0469*50) = 36.13075, rounds to 36

(Remember that for the cut score we only use the lower half of the confidence interval, and use the one-tailed value, 1.65.)

$\text{CI}_{.95}$, X, 37 = 37 ± 1.96*(.0469*50) = 32.4038-41.5962, rounds to 32-42

3. Calculate $\text{SE}_{\text{meas}(X_i)}$ and .95 CIs for scores of 47, 43, 40 and 37, where 40 is the cut score.

X = 47: $\text{SE}_{\text{meas}(47)} = \sqrt{\frac{47(50-47)}{50-1}} = \sqrt{\frac{141}{49}} = 1.696$

$\text{CI}_{.95}$ = 47 ± 1.96*1.696 = 43.67584-50.32416, rounds to 44-50

X = 43: $\text{SE}_{\text{meas}(43)} = \sqrt{\frac{43(50-43)}{50-1}} = \text{SE}_{\text{meas}(43)} = \sqrt{\frac{301}{49}} = 2.478$

$\text{CI}_{.95}$ = 43 ± 1.96*2.478 = 38.14312-47.85688, rounds to 38-48

X(λ) = 40: $SE_{meas(40)} = \sqrt{\frac{40(50-40)}{50-1}} = SE_{meas(40)} = \sqrt{\frac{400}{49}} = 2.857$

$CI_{.95}$ = 40 - 1.65*2.8577 = 35.284895, rounds to 35

(Remember that for the cut score we only use the lower half of the confidence interval, and the one-tailed value, 1.65.)

X = 37: $SE_{meas(37)} = \sqrt{\frac{37(50-37)}{50-1}} = SE_{meas(40)} = \sqrt{\frac{481}{49}} = 3.133$

$CI_{.95}$ = 37 ± 1.96*3.133 = 30.85932-43.14068, rounds to 31-43

4. Calculate Φ_{λ} for proportion cut scores of 94%, 85%, 80% and 74%.

λ= .94: $$\Phi_{.94} = 1 - \frac{1}{50-1}\left(\frac{.872(1-.872)-.00397}{(.872-.94)^2+.00397}\right) = 1 - \frac{1}{49}\left(\frac{.111616-.00397}{.004624+.00397}\right) =$$

$$1-.020408\left(\frac{.107646}{.008594}\right) = 1\text{-}.2531197 = .747$$

λ= .85: $$\Phi_{.85} = 1 - \frac{1}{50-1}\left(\frac{.872(1-.872)-.00397}{(.872-.85)^2+.00397}\right) = .507$$

λ= .80: $$\Phi_{.80} = 1 - \frac{1}{50-1}\left(\frac{.872(1-.872)-.00397}{(.872-.80)^2+.00397}\right) = .760$$

λ= .74: $$\Phi_{.74} = 1 - \frac{1}{50-1}\left(\frac{.872(1-.872)-.00397}{(.872-.74)^2+.00397}\right) = .897$$

5. Calculate $\kappa^2_{(X,T_X)}$ for cut scores of 47, 43, 40 and 37, using KR21 as an estimate of internal consistency.

λ= 47: $$\kappa^2_{(X,T_X)} = \frac{.447*9.9351+(43.6-47)^2}{9.9351+(43.6-47)^2} = \frac{4.4409897+11.56}{9.9351+11.56} =$$

$$\frac{16.0009897}{21.4951} = .744$$

λ= 43: $$\kappa^2_{(X,T_X)} = \frac{.447*9.9351+(43.6-43)^2}{9.9351+(43.6-43)^2} = \frac{4.8009897}{10.2951} = .467$$

$$\lambda = 40: \quad \kappa^2_{(X,T_X)} = \frac{.447 * 9.9351 + (43.6-40)^2}{9.9351 + (43.6-40)^2} = \frac{17.4009897}{22.8951} = .760$$

$$\lambda = 37: \quad \kappa^2_{(X,T_X)} = \frac{.447 * 9.9351 + (43.6-37)^2}{9.9351 + (43.6-37)^2} = \frac{48.0009897}{53.4951} = .897$$

Summary table

Estimates	**Score** 47 (94%)	**Score** 43 (85%)	**Score** 40 (80%)	**Score** 37 (74%)
SEM_{abs}	.0469	.0469	.0469	.0469
$\mathbf{CI}_{.95}$	42 - 52	38 - 48	36	32 - 42
$\text{SE}_{\text{meas}(X_i)}$	1.696	2.478	2.857	3.133
$\mathbf{CI}_{.95}$	44 - 50	38 - 48	35	31 - 43
$\boldsymbol{\Phi}_{\lambda}$	.747	.507	.760	.897
$\kappa^2_{(X,T_X)}$	.744	.467	.760	.897

B. Looking at the values you have calculated for the agreement indices, answer the following questions:

1. The values for the $\text{SE}_{\text{meas}(X_i)}$ get larger as the cut scores get smaller.

2. The values for Φ_λ and $\kappa^2_{(X,T_X)}$ get larger as the cut score differs from the mean.

3. If we wanted to minimize false positives, we would set the cut score at 47. This is where the $\text{SE}_{\text{meas}(X_i)}$ and its $\text{CI}_{.95}$ are the smallest, and the values for Φ_λ and $\kappa^2_{(X,T_X)}$ are acceptable. If we wanted to minimize false negatives, we would set the cut score at 37, which is where the values for Φ_λ and $\kappa^2_{(X,T_X)}$ are the largest, even though this is where the value for the $\text{SE}_{\text{meas}(X_i)}$ is the largest.

4. The nature of the domain and what test users feel is a fair and appropriate level of mastery.

C. Using the data below, for Tests 1 and 2 (X_1 and X_2), calculate the following agreement indices.

1. Calculate $\ddot{p}_o$ for cut scores of 47, 43, 40 and 37.

2. Calculate $\ddot{\kappa}$ for cut scores of 47, 43, 40 and 37.

Statistics for Tests 1 and 2 (X_1 and X_2)

	X_1	X_2
$\overline{X}$	43.6	43.4
S	3.07	3.31
r_{12}	.899	

S	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
λ	47		43		40		37	
1.	47	46	47	46	47	46	47	46
2.	47	47	47	47	47	47	47	47
3.	46	47	46	47	46	47	46	47
4.	46	46	46	46	46	46	46	46
5.	46	46	46	46	46	46	46	46
6.	46	45	46	45	46	45	46	45
7.	46	46	46	46	46	46	46	46
8	45	46	45	46	45	46	45	46
9.	45	44	45	44	45	44	45	44
10.	45	45	45	45	45	45	45	45
11.	45	46	45	46	45	46	45	46
12.	44	42	44	42	44	42	44	42
13.	44	44	44	44	44	44	44	44
14.	44	45	44	45	44	45	44	45
15.	42	38	42	38	42	38	42	38
16.	41	38	41	38	41	38	41	38

17.	40	40	40	40	40	40	40	40
18.	38	40	38	40	38	40	38	40
19.	38	37	38	37	38	37	38	37
20.	37	39	37	39	37	39	37	39
A*	1		13		15		20	
B*	1		1		2		0	
C*	1		0		1		0	
D*	17		6		2		0	

Test 2

Cut score = λ		Master	Non-master	Marginals
Test 1	Master	A	B	A+B
	Non-master	C	D	C+D
	Marginals	A+B	B+D	A+B+C+D

Test 2

Cut score = 47		Master	Non-master	Marginals
Test 1	Master	1	1	1+1=2
	Non-master	1	17	1+17=18
	Marginals	1+1=2	1+17=18	1+1+1+17=20

Test 2

Cut score = 43		Master	Non-master	Marginals
Test 1	Master	13	1	14
	Non-master	0	6	6
	Marginals	13	7	20

Test 2

Cut score = 40		Master	Non-master	Marginals
Test 1	Master	15	2	17
	Non-master	1	2	3
	Marginals	16	4	20

Test 2

Cut score = 37		Master	Non-master	Marginals
Test 1	Master	20	0	20
	Non-master	0	0	0
	Marginals	20	0	20

Equation 6.6 $\hat{p}_o = \frac{n_m}{N} + \frac{n_n}{N}$ or $\ddot{\mathbf{p}}_o = \frac{n_m + n_n}{N} = \frac{A+D}{N}$

For λ = 47: $\ddot{\mathbf{p}}_{o,47} = \frac{1+17}{20} = .900$

For λ = 43: $\ddot{\mathbf{p}}_{o,43} = \frac{13+6}{20} = .950$

For λ = 40: $\ddot{\mathbf{p}}_{o,40} = \frac{15+2}{20} = .850$

For λ = 37: $\ddot{\mathbf{p}}_{o,37} = \frac{20+0}{20} = 1.00$

Equation 6.7 $\ddot{\mathbf{p}}_c = \frac{(A+B)(A+C)+(C+D)(B+D)}{N^2}$

For λ= 47: $\ddot{\mathbf{p}}_{c,47} = \frac{(1+1)(1+1)+(1+17)(1+17)}{(20)^2} = \frac{4+324}{400} = .820$

For λ= 43: $\ddot{\mathbf{p}}_{c,43} = \frac{(13+1)(13+0)+(0+6)(1+6)}{(20)^2} = \frac{182+42}{400} = .560$

For λ= 40: $\ddot{\mathbf{p}}_{c,40} = \frac{(15+2)(15+1)+(1+2)(2+2)}{(20)^2} = \frac{272+12}{400} = .710$

For λ= 37: $\ddot{p}_{c,37} = \dfrac{(20+0)(20+0)+(0+0)(0+0)}{(20)^2} = \dfrac{400+0}{400} = 1.00$

Equation 6.7 $\ddot{\kappa} = \dfrac{(\ddot{p}_o - \ddot{p}_c)}{(1-\ddot{p}_c)}$

For λ= 47: $\ddot{\kappa} = \dfrac{(.900-.820)}{(1-.820)} = \dfrac{.080}{.180} = .444$

For λ= 43: $\ddot{\kappa} = \dfrac{(.950-.560)}{(1-.560)} = \dfrac{.390}{.440} = .886$

For λ= 40: $\ddot{\kappa} = \dfrac{(.850-.710)}{(1-.710)} = \dfrac{.140}{.290} = .483$

For λ= 37: $\ddot{\kappa} = \dfrac{(1.000-1.000)}{(1-1.000)} = \dfrac{.000}{.000} = 1.000$

Summary table

	Cut score			
Agreement index	**47**	**43**	**40**	**37**
$\ddot{p}_o$	.900	.950	.850	1.000
$\ddot{\kappa}$	.444	.886	.483	1.000

D. Looking at the values you have calculated for the agreement indices, answer the following questions:

1. The value for $\ddot{p}_o$ is highest at the mean and six points below the mean, so there doesn’t seem to be a consistent pattern.

2. The value for $\ddot{\kappa}$ is highest near the means of the two tests.

V. Illustrative research study

A. This finding reflects the data in the summary table above. The agreement indices for cut scores closer to the total group mean (67.84) are generally the lowest, while those that are further from the total group mean are larger.

B. Both the Φ coefficient and KR-20 look at the same source of error – inconsistencies across items. The other indices take the cut score into consideration, albeit in different ways

C. With the exception of the cut score for 33A, all the cut scores are below the group means. As with the data above, the $\hat{\kappa}$ indices are much lower than the $\hat{p}_o$ indices, especially for the lower scores. As would be expected, the values for Φ_λ are higher as scores are further from the total group mean.

D. The values for coefficient Φ and KR-20 are much larger for the total than for any of the groups because these are based on a larger sample than the individual groups.

E. In this situation, we think the most important information is the dependability of placement decisions, so we would use an agreement index. Coefficient Φ and KR-20 do not take the cut score into account, so provide no information about decision consistency. In this situation, we think it is also important to treat classification errors far from the mean of the test as more serious, so would use a squared-error-loss agreement index. We think the Φ_λ is the most appropriate, because this provides information about the dependability of the classification decisions at the different cut scores, and treats misclassifications far from the cut score as more serious than those near the cut score. The $\hat{p}_o$ and $\hat{\kappa}$ indices treat all misclassifications as equally serious, so would not be appropriate for this situation.