

**Texas State University
Testing, Research-Support, and Evaluation Center**

How to Read Faculty Exam Output

Mis-Keyed Items

The preface page of the faculty exam output may contain a list of possible mis-keyed items (i.e., items may be incorrectly bubbled on the key sheet) and their associated form (if applicable). In the example, item 21 on form 2 is flagged as a possible mis-keyed item. The faculty member should review the item analysis (discussed later) to determine if the wrong response choice has been bubbled. If the response choice should be changed, the faculty member needs to call TREC to request that the keyed answer be changed and the exam output reprocessed (TREC will edit the data in the scan file, but will not alter any marks on the hard copy of the key sheet).

***** POSSIBLE MIS-KEYED FORM(S)! *****

The following Items may be mis-keyed. Items listed below have a negative Item Discrimination and an Item Difficulty below 25% and there was another choice for the item which was selected by more than 50% of the students. Items which meet these criteria have been found to be mis-keyed more than 80% of the time. Please check the Item Analysis for these Items.

Key Form#	Question#
-----	-----
02	21

Students Scored with Wrong Key

The preface page of the faculty exam output may also contain a list of students who have possibly been scored with the wrong key. This section prints only for exams that have multiple forms. In the example, the student scored below a 30% on an exam with two forms. The student's raw score using key 1 was 15 and the student's raw score using key 2 was 44. The sheet# indicates the location of the scan sheet. In this case, the faculty member needs to determine if the student's exam was scored using the wrong key (TREC will take no further action).

*** Student(s) Possibly Scored With Wrong Key! ***

The following student(s) have scored below 30%. They are checked against the other key(s).
The sheet# will help you to locate the student's sheet within the stack of scan sheets.

ID#	NAME	KEY	SCORE	BETTER KEY	BETTER SCORE	SHEET#
-----	-----	-----	-----	-----	-----	-----
A88888888	DOE JANE C	01	15	02	44	10

If there are no items to be flagged as possibly mis-keyed, and no students to be flagged as possibly scored with the wrong key, there will be no preface page, and your output will begin with your roster.

Rosters

There are two roster types that may be selected for output: Alphabetical and Numeric. An Alphabetical roster lists students by last name, first name, middle initial, and student ID number, separated for each section submitted. Both roster types will also print a form number for each student. This is the form number of the key against which the student's responses were scored. This is then followed by the raw number of questions answered correctly, and the percentage correct.

ALPHABETICAL ROSTER							
For:		18 Students		Class\Section: 4301 251			
Processed on: 08-10		at 08:59:04		File Name: C:\TREC\DATA\TESTDATA.SDF			
LAST NAME	FIRST	MI	ID#	FORM#	#CORRECT	%CORRECT	
(none)	012*		A66666666	02	23 92
BLOW	JOE	X	A11111111	02	21 84
PEEP	BO	L	A22222222	02	18 72

In the example above, the first line has no student name. This means the student did not bubble in his/her name on the answer sheet. The number 012 to the right of (none) indicates the location of the scan sheet in the stack. In this case, it is the 12th sheet from the top. In some cases (especially during finals) the exam might be scanned with the Opscan 8 (the second scanner). In this case the number 012 indicates that the exam is the 12th sheet from the bottom.

The Numeric roster omits all name data, and lists students by ID numbers only.

NUMERIC ROSTER					
For:		18 Students		Class\Section: 4301 251	
Processed on: 08-10		at 08:59:04		File Name: C:\TREC\DATA\TESTDATA.SDF	
ID#	FORM#	#CORRECT	%CORRECT		
A33333333	02	21	84
A77777777	02	23	92
(none)	02	19	76

If your exam has a point maximum other than 100, an additional "points" column will be added to the roster. In the example below, the exam had 25 questions and was worth 85 points, so each question would be worth 3.4 points.

ALPHABETICAL ROSTER								
For:		18 Students		Class\Section: 4301 251				
Processed on: 08-10		at 13:12:03		File Name: C:\TREC\DATA\TESTDATA.SDF				
LAST NAME	FIRST	MI	ID#	FORM#	#CORRECT	%CORRECT	POINTS (85 MAX)	
BLOW	JOE	X	A11111111	02	21 84 71.4
PEEP	BO	L	A22222222	02	18 72 61.2

Another possible variation is for exams with weighted questions. If some questions on your exam are worth a different number of points than other questions, you will need to use the weighted option. In this case, the output file will provide raw number correct, points according to your assigned weights, and then a weighted percentage score:

ALPHABETICAL ROSTER								
For:		48		Students		Class\Section: 3333 001		
Processed on:		02-16		at 13:36:27		File Name: C:\TREC\DATA\TESTDATA.SDF		
LAST NAME	FIRST	MI	ID#	FORM#	#CORRECT	WEIGHTED POINTS	WEIGHTED %CORRECT	
BUNNY	BUGS	C	A55555555	02 41 10384	
DUCK	DAFFY	R	A44444444	01 29 53 52	
FUDD	ELMER	D	A99999999	02 25 53 52	
YOSEMITE	SAM	X	A00000000	01 41 10082	

In the example above, there were 51 questions, with a maximum possible score of 122 points (31 questions worth 2 points each, and 20 questions worth 3 points each). Notice that the first and fourth students in the example above both got the same number of correct answers, but it has translated to different numbers of points because of the weighted scoring. Similarly, the second and third students got different numbers of questions correct, but achieved the same score. The final column (weighted percent) is calculated by dividing the weighted points by the maximum possible points.

Frequency Distributions

The frequency distribution shows both numerically and graphically the scores which the students received, and how many students received each score. In the example below, the first column shows the raw scores which the students received. The raw scores ranged from 13 to 23. The second column shows the equivalent percentage correct for each raw score. The third column, labeled *Frequency* shows how many students received each raw score. In the example, 1 student received a raw score of 13 and 5 students received a raw score of 20. The next column is a running total of the frequencies so the final line will show the total number of students in the distribution. The *Percent* column shows the percentage of students who earned each particular score. In the example, you can see 5 students answered 20 questions correctly. There are 18 students in the distribution. Therefore, 5/18 of the students or 27.8% had 20 correct. The next column is a running total of the percent column and must end with 100.0. The final column is a duplicate of the first column (raw score) and is used for reference. To the right of the last column is a graphical representation of the score distribution. Each dash represents one student. Frequency distributions can be generated one for each *section*, one for each *form*, or one for the entire run. Any one, or all three may be chosen.

FREQUENCY DISTRIBUTION OF SCORES BY SECTION						
Processed on: 08-10 at 08:59:04 File Name: C:\TREC\DATA\TESTDATA.SDF						
Class: 4301 Section: 251 Students: 18						
Score		Cumulative		Cumulative		
Raw (%)	Frequency	Frequency	Percent	Percent	Score	
13 (52.0)	1	1	5.6	5.6	13_	
14 (56.0)	1	2	5.6	11.1	14_	
15 (60.0)	1	3	5.6	16.7	15_	
18 (72.0)	4	7	22.2	38.9	18____	
19 (76.0)	1	8	5.6	44.4	19_	
20 (80.0)	5	13	27.8	72.2	20_____	
21 (84.0)	2	15	11.1	83.3	21__	
23 (92.0)	3	18	16.7	100.0	23___	
Each '_' represents 1 Student						
Mean Score	:19.1	(76.4%)				
Median Score	:20.0	(80.0%)				

If weighted scores are used, the first and last columns will be *Weighted Points* rather than raw score, and the second column will be the weighted percentage.

Mean/Median

The mean is the most commonly used estimate of a typical score. It is the sum of all the scores (per distribution), divided by the number of students. The median is the score which splits the distribution in half so that half the students score above and half score below it. In the example above, the mean is 19.1 and the median falls at a score of 20. The mean and the median will be similar if the scores are symmetrically distributed. However, when there are some extreme scores, the median may be a better estimate of central tendency.

Item Analysis

When scoring an exam, an Item Analysis will be generated for each *unique* key form used. In the example below, you can see that there were 25 items on key form #02. The first column of the Item Analysis lists the item numbers. The next 5 columns are titled A, B, C, D, & E. These refer to the 5 choices each question had (you can have up to 10 response choices, depending on the type of key and answer sheet used). For item 1, no student chose D or E, while 2 students chose A, 1 student chose B, and 15 students chose C. The asterisk indicates the correct answer; for item #1, C is the correct answer. If you choose multiple answers correct or credit for all, all correct responses will have asterisks. The next column indicates how many students left the item blank.

ITEM ANALYSIS									
Processed on 08-10		at 08:59:04		File Name: C:\TREC\DATA\TESTDATA.SDF					
18 students		Test Form #: 02		25 Items					
Item#	Responses					Blank	Item Difficulty	Item Discrim.	Item#
	A	B	C	D	E				
1	2	1	15*	0	0	0	83.3	.2113	1
2	3	14*	1	0	0	0	77.8	.1143	2
3	3	0	5	10*	0	0	55.6	-.1403	3
4	11	7*	0	0	0	0	38.9	.2116	4
5	16*	2	0	0	0	0	88.9	.2911	5
6	1	0	17*	0	0	0	94.4	.0145	6
7	0	0	18*	0	0	0	100.0	.0000	7
8	15*	2	1	0	0	0	83.3	-.0089	8
9	15*	0	3	0	0	0	83.3	.0996	9
10	0	16*	1	1	0	0	88.9	.6436	10
11	2	0	16*	0	0	0	88.9	-.2199	11
12	1	14*	3	0	0	0	77.8	.1645	12
13	16*	0	2	0	0	0	88.9	.2239	13
14	2	1	15*	0	0	0	83.3	.7615	14
15	2	3	13*	0	0	0	72.2	.5880	15
16	5	1	12*	0	0	0	66.7	.0285	16
17	14*	4	0	0	0	0	77.8	.0648	17
18	12*	4	2	0	0	0	66.7	.3948	18
19	2	15*	1	0	0	0	83.3	.0996	19
20	2	2	14*	0	0	0	77.8	.5393	20
21	15	2*	0	1	0	0	11.1	-.3650	21
22	1	15*	1	1	0	0	83.3	.0449	22
23	1	5	12*	0	0	0	66.7	.5453	23
24	1	0	17*	0	0	0	94.4	.0145	24
25	14*	3	1	0	0	0	77.8	-.2175	25

(For Responses, an '*' = Key)

DISTRIBUTION OF ANSWERS ON KEY:

A	B	C	D	E
7	7	10	1	0

OF CASES MEAN STANDARD DEVIATION STANDARD ERROR

18 19.11 (76.44%) 2.81 1.86

RELIABILITY: .559

Item Difficulty

The item difficulty column shows the percentage of students who answered each item correctly. Everyone answered item 7 correctly, so the item difficulty for this item is 100.0.

Item Discrimination

The item discrimination is a correlation between the points awarded on an item and the total test score. When the item discrimination is *positive*, students who answered the item correctly performed better on the rest of the test than students who answered the item incorrectly. When the item discrimination is *negative*, students who answered the item incorrectly did better on the rest of the test than students who answered the item correctly. Item discriminations of 0 mean that there is no difference between the two groups. In the sample item analysis, item 7 has an item discrimination of 0. Everyone answered this item correctly, so there were not two groups to discriminate between (i.e., those who answered correctly vs. those who answered incorrectly). Item 8 shows that 15 students answered the item correctly and 3 answered the item incorrectly. The item discrimination is slightly negative for this item, which means that the 3 students who

missed the item actually did better on the test than the 15 students who answered correctly. This can often be interpreted in three ways. The item is:

- a. dissimilar from the other items, not measuring or assessing what the other items are measuring or assessing;
- b. ambiguous or poorly worded; or
- c. mis-keyed (the wrong answer was bubbled on the answer key).

In the example above, item 21 was intentionally mis-keyed. As this item illustrates, there is a very low passing rate (i.e., low item difficulty) and negative item discrimination. If an item is determined to be poorly worded or ambiguous, it can be dropped from the exam by leaving the answer key blank for that item. If the faculty member decides this is the case, he/she can contact the Testing, Research-Support, and Evaluation Center at 5-2276 and request that the item be dropped. The center makes the requested changes and reruns the rosters.

Standard Deviation

The standard deviation is a measure of how far the scores deviate from the mean. The more spread out the scores the larger the standard deviation.

Standard Error

The concept of the standard error of measurement involves repeated testing of an individual with the test. Because of slight differences in testing conditions and individual responses, the scores on the test will likely not be the same. Instead there will be a distribution of scores, and the mean of the distribution is generally considered to be the best estimate of the person's ability. The standard error of measurement is the standard deviation of this distribution. The value computed by the scoring program uses information from only one administration of a test. By assuming that the distribution of scores would be normally distributed, the standard error of measurement can be estimated. In the example, the SEM is 1.86. This may be interpreted to mean that the score which best represents a student's true capability will be within 1.86 points of his or her raw score about 68% of the time, or within 3.72 points 95% of the time.

Reliability

This coefficient is an estimate of the extent to which each item measures what the entire test is measuring and can be thought of as an index of a test's internal consistency. Reliability will have a value between 0.00 and 1.00. A reliability of 1.00 means that each item measures exactly the same ability as does the total test score, while a reliability of 0.00 means that the item scores are unrelated to the total test. Values between approximately .60 and .80 are typical for classroom tests.

Reliability is influenced by several test characteristics. Most important is similarity of the content of the items. A test which attempts to assess many different abilities will probably have lower reliability than a test which attempts to measure only one ability. Also of importance is the length of the test. In general, longer tests are more reliable than shorter tests. A third influence on reliability is difficulty. Tests which are very easy or very difficult tend to be less reliable than those of moderate difficulty.

The reliability of a test may be increased by using the information provided in the *Item Analysis*. The most important thing to look at is the *Item Discrimination*. A test's reliability will be increased by deleting items with negative discriminations (even though the test will be shortened). However, when considering deleting any item from a test, remember that the item statistics should only serve as a guide and that small increases in reliability are less important than representative coverage of the topic area. This is particularly true when class size is below 30.

Item Removal Analysis

If there are any items with negative discriminations on a test, then an Item Removal Analysis will print at the end of the output. Statistics will be regenerated allowing you to see how dropping an item would affect the overall statistics (presuming all items with lower discriminations have also been removed). In the example, you can see that item 21 had a negative discrimination of $-.365$. If this item were dropped from the test, then the new mean score would be 19, the new maximum score would be 23, the new minimum score would be 13, the new standard deviation (SD) would be 2.91, the new standard error (SEM) would be 1.83, and the new reliability would be $.605$. As this example illustrates, dropping all items with negative discriminations would yield a reliability of $.712$. This output is included with your results but items are only dropped if you call TREC to indicate that we should drop them.

ITEM REMOVAL ANALYSIS								
Effects of Removing Questions with Negative Item Discriminations:								
Ques#	Discr.	#Left	Mean	Max	Min	SD	SEM	Rel.
21	-0.365	24	19.00 (79.17)	23 (95.83)	13 (54.17)	2.91	1.83	.605
11	-0.220	23	18.11 (78.74)	22 (95.65)	12 (52.17)	2.94	1.79	.629
25	-0.217	22	17.33 (78.79)	21 (95.45)	11 (50.00)	3.00	1.73	.666
3	-0.140	21	16.78 (79.89)	21 (100.00)	10 (47.62)	2.99	1.65	.694
8	-0.009	20	15.94 (79.72)	20 (100.00)	9 (45.00)	2.99	1.60	.712