8 Analyzing Classroom Assessment Data

Brian C. Wesolowski

Chapter Overview

This chapter focuses on how to describe and analyze classroom testing data, with the intent of informing future teaching and learning processes and improving future test uses from both a class-centered perspective and an individual student-centered perspective. Concepts covered include item- and person-ordering, item difficulty, person ability, item- and person-discrimination, and distractor analyses.

Learning Expectations for the Chapter

- Examine the relationship between person ability and item difficulty.
- Calculate and evaluate item difficulty and person ability.
- Calculate and evaluate item- and person-discrimination.
- Use distractor analyses to better understand abnormalities in the outcome testing data.

Essential Questions for the Chapter

- How can learning outcome data be used in a way that informs teaching and learning while also communicating to administrators the types and quality of teaching and learning occurring in the music classroom?
- How do I calculate item difficulty and item-discrimination indices?
- How can item difficulty and item-discrimination indices inform class-centered learning outcomes?
- How do I calculate person ability and person-discrimination indices?
- How can person ability and person-discrimination indices inform student-centered learning outcomes?
- How do distractor analyses provide more meaningful insight into response patterns?

Today's educational environment is becoming increasingly data-driven, and there is a clear need to communicate to administrators and other educational stakeholders the teaching and learning occurring in music classrooms using empirical data (Wesolowski, 2014, 2015). As discussed in Chapter 5, the literature pertaining to the procedural and analytical methods for demonstrating student achievement is most often in the context of large-scale, standardized tests. However, 69% of classroom educators, including music educators, instruct students in a discipline where they are not evaluated in the context of standardized testing (National Comprehensive Center for Teacher Quality, 2011). For these 69% of educators, it becomes their responsibility to communicate to stakeholders the representative student achievement in their classrooms using classroom assessments. From a music education perspective, this can be even more daunting, as administrators and stakeholders are not often familiar with teaching and learning processes specific to the field of music (Hart, 2003). Therefore, it is critical for music educators to provide data from assessments that validly, reliably, and fairly represent the true teaching and learning within the music classroom in a way that administrators and stakeholders can understand. This chapter reviews procedures to generate, collect, and disseminate data in a way that can provide empirical evidence of music student achievement in meaningful ways while generating empirical evidence to make informed inferences related to the quality of the test itself.

Setting the Stage

Let us suppose that a music educator teaches a unit to their general music class on instrument timbre. One part of the teacher's overall assessment is to provide a multiple-choice listening test with 20 samples of music, highlighting various performances of instruments discussed in the unit. For each musical sample played, the students are asked to select which musical instrument is performing from one of four choices using a sound-to-picture multiple-choice format. As an example, for Item 1, the student would be prompted with pictures of a piano, clarinet, trumpet, and timpani. The teacher would play an example of a solo piano sonata, and the student would answer the item correctly by circling a picture of the piano.

The teacher will rely on the results of the test as one part of the overall student assessment in order to observe how successful the instructional delivery was and to identify how well the students have learned to aurally identify various instrument timbres. Based upon what the teacher knows about the students from the class interactions, some musical samples were selected because they were considered relatively easy to identify and it is anticipated that most students would correctly identify the instrument. Other musical samples were selected because they were considered to

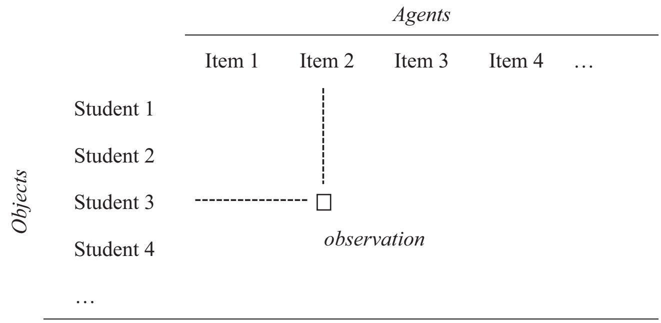

Figure 8.1 A representation of an observation stemming from the interaction between one object (Student 3) and one agent (Item 2)

be relatively difficult and only students possessing higher levels of aural discrimination would be able to correctly identify the instrument. Other musical selections sit somewhere in the middle, where it is anticipated that that some students would correctly identify the instrument while others would not.

There are two important outcomes that can be drawn from this testing context. The first piece of information is the expected ordering of the students from highest ability to the lowest ability. In this instance, the students are considered to be the *objects* of the testing context because they are being evaluated. The second piece of information is the expected ordering of the test items from the most difficult to the least difficult. In this instance, the items are considered to be the *agents* of the testing context because they are doing the evaluating. Throughout the testing process, each object (i.e., student) interacts with each agent (i.e., item). Each of these interactions is referred to as a raw score response, or an *observation* (see Figure 8.1). The considerations toward the ordering of objects, the ordering of agents, and all the individual observations provide a teacher with the overall picture of the testing context. It is the music educator's job to then evaluate, interpret, and diagnose the truthfulness of the outcomes. The teacher needs to ask, *Do the outcomes cooperate with my intentions and expectations behind the test while also representing the true teaching and learning occurring in the classroom?*

Constructing a Data Matrix and Right/Wrong Matrix

The observations for the hypothetical example described earlier can be found in the data matrix depicted in Table 8.1. In this example, 18 students responded to 20 multiple-choice items, resulting in a total of 360

Table 8.1 Data matrix containing raw score responses of 18 students to 21 items

											Agents										
	Items	*I01*	*I02*	*I03*	*I04*	*I05*	*I06*	*I07*	*I08*	*I09*	*I10*	*I11*	*I12*	*I13*	*I14*	*I15*	*I16*	*I17*	*I18*	*I19*	*I20*
	Correct Answers	*3*	*1*	*1*	*1*	*4*	*1*	*3*	*3*	*1*	*3*	*2*	*1*	*1*	*2*	*3*	*2*	*1*	*1*	*2*	*2*
Objects	**Student 1**	1	3	2	3	4	2	2	1	2	3	2	3	2	1	3	2	3	1	1	1
	Student 2	3	1	3	3	2	1	2	1	1	2	2	1	1	2	1	1	1	1	2	2
	Student 3	4	2	2	2	3	2	3	2	1	2	1	1	1	2	2	1	2	1	2	2
	Student 4	3	1	1	1	4	1	3	3	1	3	2	1	1	2	1	2	1	1	2	2
	Student 5	3	1	1	1	4	1	2	3	1	3	2	2	2	2	2	2	1	1	2	2
	Student 6	1	2	3	1	2	1	2	3	1	3	2	3	1	1	3	2	3	1	1	2
	Student 7	1	3	3	1	2	2	1	3	2	2	1	2	1	2	1	1	3	2	2	2
	Student 8	1	2	3	3	2	3	1	1	1	2	2	2	2	2	1	2	3	1	2	1
	Student 9	1	3	1	1	4	1	2	3	1	3	2	2	2	2	1	2	3	1	2	2
	Student 10	3	1	3	1	2	1	3	1	1	3	2	2	1	2	2	3	3	1	2	2
	Student 11	3	1	1	1	4	1	3	1	1	3	3	1	1	2	1	2	3	1	2	2
	Student 12	3	3	1	3	1	3	2	1	1	3	1	1	2	1	2	2	1	2	1	2
	Student 13	3	3	1	1	4	1	2	3	1	3	2	1	2	2	3	2	3	1	2	2
	Student 14	3	3	1	1	2	1	2	1	1	3	2	2	1	1	1	3	3	1	2	2
	Student 15	3	1	3	1	4	1	3	3	1	3	2	1	1	2	2	2	1	1	2	2
	Student 16	2	1	2	2	4	2	3	2	2	3	2	2	1	1	3	3	1	2	2	1
	Student 17	3	1	1	1	4	1	3	3	1	3	2	2	2	2	2	2	3	1	2	2
	Student 18	1	3	3	1	4	1	2	1	1	3	2	1	1	2	1	1	3	1	2	2

observations (18 students multiplied by 20 items). Here, the students are ordered in some type of pre-established order from top to bottom, either alphabetically, or by student number if the teacher wishes to keep the results anonymous. The items are ordered in the order in which they appear on the test. The answer key is provided in the second row along with person[1] responses to each item. In this example, the data are coded to where choice A = 1, choice B = 2, choice C = 3, and choice D = 4. Coding the responses as numerical input is important, as the analyses described in the chapter will be empirical in nature.

Once the observations are compiled into a data matrix, the next step is to convert the data matrix to a right/wrong matrix (see Table 8.2). For most multiple-choice tests, there is one correct answer and all other response options are incorrect. In these instances, every observation is either correct or incorrect. When the observations result in either a correct or incorrect response, the test responses are considered to be **dichotomous**. The right/wrong matrix is a representation of the dichotomous responses, coded as either 0 = incorrect (i.e., "wrong") or 1 = correct ("right"). The right/wrong matrix will form the foundation for much of the remaining data analysis processes.

Within the right/wrong matrix, the students can be ordered from highest ability to lowest ability based upon the *person sum score* (the total of correct answers for each person). Additionally, the items can be ordered from the least difficult to the most difficult based upon the *item sum score* (the total of correct answers for each item) (see Table 8.3). In creating these orderings, an interesting pattern emerges. If a diagonal line were to be drawn from the top right part of the matrix down to the bottom left part of the matrix, we would see that above the line, more 1s would emerge, particularly as the observations approach the top left part of the matrix (closest to the observation represented by the interaction between the highest ability student and least difficult item – Student 4 and Item 19) (see Figure 8.2). This indicates that the higher the person ability, the more likely the person is to respond correctly to a less difficult item. Oppositely, below the line, more 0s emerge, particularly as the observations approach the bottom right part of the matrix (closest to the observation represented by the interaction between the lowest ability student and most difficult item – Student 1 and Item 15. This indicates that the lower the person ability, the more likely the person is to respond incorrectly to a more difficult item. The closer that the observations approach the diagonal line, the more inconsistencies are observed, indicating more randomness to the response patterns.

The visual elements in the right/wrong matrix with the ordering of items and persons as described above provide more apparent insights into some inconsistencies in the response patterns. From an item perspective,

Table 8.2 Right/wrong matrix consisting of 360 dichotomously scored observations

		Agents																			
		I01	*I02*	*I03*	*I04*	*I05*	*I06*	*I07*	*I08*	*I09*	*I10*	*I11*	*I12*	*I13*	*I14*	*I15*	*I16*	*I17*	*I18*	*I19*	*I20*
Objects	**Student 1**	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	0	1	0	0
	Student 2	1	1	0	0	0	1	0	0	1	0	1	1	1	1	0	0	1	1	1	1
	Student 3	0	0	0	0	0	0	1	0	1	0	0	1	1	1	0	0	0	1	1	1
	Student 4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
	Student 5	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	1	1	1	1	1
	Student 6	0	0	0	1	0	1	0	1	1	1	1	0	1	0	1	1	0	1	0	1
	Student 7	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	0	1	1
	Student 8	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	1	1	0
	Student 9	0	0	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	1	1
	Student 10	1	1	0	1	0	1	1	0	1	1	1	0	1	1	0	0	0	1	1	1
	Student 11	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	0	1	1	1
	Student 12	1	0	1	0	0	0	0	0	1	1	0	1	0	0	0	1	1	0	0	1
	Student 13	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1
	Student 14	1	0	1	1	0	1	0	0	1	1	1	0	1	0	0	0	0	1	1	1
	Student 15	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
	Student 16	0	1	0	0	1	0	1	0	0	1	1	0	1	0	1	0	1	0	1	0
	Student 17	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1
	Student 18	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0	0	0	1	1	1

Table 8.3 Right/wrong matrix with ordering of students from high ability to low ability and ordering of items from least difficult to most difficult

		Agents																				*Person Sum Score*
		I19	*I20*	*I18*	*I10*	*I11*	*I09*	*I14*	*I04*	*I06*	*I13*	*I16*	*I01*	*I05*	*I02*	*I03*	*I08*	*I12*	*I07*	*I17*	*I15*	
Objects	**Student 4**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	**19**
	Student 15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	**18**
	Student 11	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	**16**
	Student 17	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0	**16**
	Student 13	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	1	**16**
	Student 5	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	**15**
	Student 10	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	1	0	0	**13**
	Student 9	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	0	0	0	0	**13**
	Student 6	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	1	**12**
	Student 18	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0	**12**
	Student 14	1	1	1	1	1	1	0	1	1	1	0	1	0	0	1	0	0	0	0	0	**11**
	Student 2	1	1	0	0	1	0	1	0	1	1	0	1	0	1	0	0	1	0	1	0	**10**
	Student 16	1	0	0	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	1	1	**9**
	Student 3	1	1	1	0	0	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0	**8**
	Student 8	1	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	**7**
	Student 12	1	1	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	**7**
	Student 7	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	**6**
	Student 1	0	0	1	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	**5**
	Item Sum Score	**17**	**16**	**14**	**14**	**14**	**13**	**13**	**12**	**12**	**11**	**11**	**10**	**10**	**8**	**8**	**8**	**8**	**7**	**4**	**3**	

	Agents																				
Objects	119	120	118	110	111	109	114	104	106	113	116	101	105	102	103	108	112	107	117	115	Person Sum Score
Student 4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	**19**
Student 15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	**18**
Student 11	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	**16**
Student 17	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0	**16**
Student 13	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	1	**16**
Student 5	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	**15**
Student 10	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	1	0	0	**13**
Student 9	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	0	0	0	0	**13**
Student 6	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	1	**12**
Student 18	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0	**12**
Student 14	1	1	1	1	1	1	0	1	1	1	0	1	0	0	1	0	0	0	0	0	**11**
Student 2	1	1	0	0	1	0	1	0	1	1	0	1	0	1	0	0	1	0	1	0	**10**
Student 16	1	0	0	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	1	1	**9**
Student 3	1	1	1	0	0	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0	**8**
Student 8	1	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	**7**
Student 12	1	1	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	**7**
Student 7	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	**6**
Student 1	0	0	1	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	**5**
Item Sum Score	**17**	**16**	**14**	**14**	**14**	**13**	**13**	**12**	**12**	**11**	**11**	**10**	**10**	**8**	**8**	**8**	**8**	**7**	**4**	**3**	

Figure 8.2 Recognizing abnormalities in response patterns in the context of item- and person-ordering

we can look down the columns for areas where unexpected responses occur. Some examples include:

- *Item 13*: There is a string of correct responses for Students 14, 2, 16, and 3 where we would expect incorrect answers and a string of incorrect responses for Students 17, 13, and 5 where we would expect correct responses.
- *Item 12*: Students 18, 2, 3, and 12 answered correctly when we would expect them to answer incorrectly.
- *Item 14*: Students 3, 8, and 7 answered correctly when we would expect them to answer incorrectly.

The unexpected patterns in the item behavior provide an initial, qualitative awareness into potentially problematic items on the test from a class perspective.

From a person perspective, we can look across the rows for areas where unexpected responses occur. Some examples include:

- Student 11: Answered Item 11 incorrectly when it was expected to be answered correctly.
- Student 2: Answered some items incorrectly that were expected to be correct (Items 20, 18) and answered several items correctly that were expected to be incorrect (Items 1, 2, 12, 17).
- Student 16: Answered several items incorrectly that were expected to be correct (Items 18, 10, and 9) and answered several items correctly that were expected to be incorrect (Items 6, 13, 1, 2, 12, and 17).

The unexpected patterns in person behavior provide an initial, qualitative awareness into errors in the test taking procedure, guessing, or an atypical understanding/interpretation of instructional content from an individual student perspective.

Empirical Investigations Into Item and Person Functioning

The remainder of the chapter provides methods for empirically investigating item-centered and person-centered data. Item-centered data, or **item functioning**, play an important role in diagnosing and evaluating class-centered behaviors with their engagement with the test items through the evaluation of item response patterns. Person-centered data, or **person functioning**, play an important role in diagnosing and evaluating individual student-centered behaviors with their engagement with the test items through the evaluation of person response patterns.

Item Difficulty Indices

Item difficulty indices are important for exploring the proportion of students who answered an item correctly and incorrectly. Item difficulty is represented as a *p*-value (proportion value) and is calculated as follows:

$$P_i = \frac{R_i}{T_i},$$

where p_i = difficulty of item *i*,
R_i = the sum of students who responded correctly to item *i*, and
T_i = the total number of students who responded to item *i*.

From our example, if we were to calculate the item difficulty for Item 10, we see from Table 8.3 that a total of 14 students out of 20 answered the item correctly. Therefore, p_{i10} would be calculated as follows:

$$P_{i10} = \frac{14}{20} = 0.70.$$

The item difficulty calculations for all 20 items can be found in Table 8.4.

The resulting values for item difficulty are a decimal ranging from 0.00 to 1.00. The closer the value approaches to 0.00, the more difficult the item is. The closer the value approaches to 1.00, the less difficult the item is. The decimal can also be interpreted as a percentage correct. In the case of Item 10, 0.70 (70%) of the students answered the item correctly. From a strict item analysis perspective,[2] item difficulty values can be interpreted as follows:

- *Easy item*: 0.75–1.00 (75%–100% of students answered the item correctly)
- *Average-difficulty item*: 0.25–0.75 (25%–75% of the students answered the item correctly)
- *Difficult item*: 0.00–0.25 (0%–25% of the students answered the item correctly)

According to Lord (1952), the ideal item difficulty in terms of discrimination potential for a five-response multiple-choice item is 0.70, the ideal item difficulty for a four-response multiple-choice item is 0.74, the ideal item difficulty for a three-response multiple-choice item is 0.77, and the ideal item difficulty for a true/false item or a two-response multiple-choice item is 0.85.

Person Ability Indices

Person ability indices are important for exploring the proportion of items that were answered correctly or incorrectly by an individual student. Person ability indices can be thought of conceptually and mathematically in

Table 8.4 Calculation of item difficulty for all 20 items

		Agents																			
		I19	*I20*	*I18*	*I10*	*I11*	*I09*	*I14*	*I04*	*I06*	*I13*	*I16*	*I01*	*I05*	*I02*	*I03*	*I08*	*I12*	*I07*	*I17*	*I15*
Objects	**Student 4**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	Student 15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0
	Student 11	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
	…	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0
	Student 1	0	0	1	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
	Item Sum Score	17	16	14	14	14	13	13	12	12	11	11	10	10	8	8	8	8	7	4	3
	p_i	0.94	0.89	0.78	0.78	0.78	0.72	0.72	0.67	0.67	0.61	0.61	0.56	0.56	0.44	0.44	0.44	0.44	0.39	0.22	0.17

the same manner as item difficulty, only from a person-centered perspective. Person ability is also represented as a *p*-value (proportion value) and is calculated as follows:

$$P_p = \frac{R_p}{T_p},$$

where p_p = ability of person p,
R_p= the sum of the items that person p responded correctly to, and
T_p = the total number of items that person p responded to.

From the example, if we were to calculate the person ability for Student 17, we see from Table 8.3 that Student 17 answered a total of 16 out of the 20 items correctly. Therefore, *pp17* would be calculated as follows:

$$P_{p17} = \frac{16}{20} = 0.80.$$

The person ability calculations for all 18 students can be found in Table 8.5.

Table 8.5 Calculation of person ability for all 18 students

		Agents					*Person Sum Score*	p_p
		I19	*I20*	*I18*	. . .	*I15*		
Objects	**Student 4**	1	1	1	. . .	0	19	0.90
	Student 15	1	1	1	. . .	0	18	0.86
	Student 11	1	1	1	. . .	0	16	0.76
	Student 17	1	1	1	. . .	0	16	0.76
	Student 13	1	1	1	. . .	1	16	0.76
	Student 5	1	1	1	. . .	0	15	0.71
	Student 10	1	1	1	. . .	0	13	0.62
	Student 9	1	1	1	. . .	0	13	0.62
	Student 6	1	1	1	. . .	1	12	0.57
	Student 18	1	1	1	. . .	0	12	0.57
	Student 14	1	1	1	. . .	0	11	0.52
	Student 2	1	1	0	. . .	0	10	0.48
	Student 16	1	0	0	. . .	1	9	0.43
	Student 3	1	1	1	. . .	0	8	0.38
	Student 8	1	1	1	. . .	0	7	0.33
	Student 12	1	1	0	. . .	0	7	0.33
	Student 7	1	1	0	. . .	0	6	0.29
	Student 1	0	0	1	. . .	0	5	0.24
	Item Sum Score	17	16	14	. . .	3		

Similarly to the interpretation of item difficulty values, person ability values result in a decimal ranging from 0.00 to 1.00. The closer the value approaches to 0.00, the lower-ability the person is. The closer the value approaches to 1.00, the higher-ability the person is. The decimal can also be interpreted as a percentage correct. In the case of Student 17, 0.80 (80%) of the items were answered correctly. From a strict person analysis perspective, person ability values can be interpreted as follows:

- *High-ability person*: 0.75–1.00 (75%–100% of items were answered correctly)
- *Average-ability person*: 0.25–0.75 (25%–75% of items were answered correctly)
- *Low-ability person*: 0.00–0.25 (0%–25% of items were answered correctly)

Item difficulty and person ability indices are proportion correct values that can be used to indicate a rank ordering of items (from least difficult to most difficult) and a rank ordering of persons (from highest-ability to lowest-ability). The problem with using only these indices, however, is that they do not provide any type of empirical quality indicator. According to Table 8.4, Items 18, 10, and 11 each have an item sum score of 14 with a *p*-value of 0.78. However, Table 8.3 indicates a different pattern of responses to each of the items. Therefore, although the *p*-values are the same, the quality of the item response patterns is different. Similarly, according to Table 8.5, Students 11, 17, and 17 each have a sum score of 16 with a *p*-value of 0.76. However, Table 8.3 indicates a different pattern of responses to each of the persons. Therefore, although the *p*-values are the same, the quality of the student response patterns is different. In order to empirically investigate the differences in quality of response patterns for both items and persons, item- and person-discrimination indices can be used.

Item-Discrimination Indices

Item-discrimination indices are important for empirically exploring the quality of the response patterns of the items. Conceptually, item-discrimination can be thought of as a value that represents the frequency with which items are responded to correctly by varying groups of ability levels of students, such as comparing high-ability student response patterns to low-ability response patterns, for example. Item discrimination is represented by a *D*-value (discrimination value) and is calculated as follows:

$$D_i = p_{i\_high} - p_{i\_low},$$

where D_i = discrimination of item *i*,
p_{i_high} = item difficulty index of the high-ability group, and
p_{i_low} = item difficulty index of the low-ability group.

In order to arrive at the calculation of D_i, there are some considerations to be made and steps to go through:

1. Start by creating a right/wrong matrix, ensuring that the persons are ordered from highest-ability to lowest-ability.
2. Divide the students evenly into a high-ability group and low-ability group. Some have suggested that the groups be divided into an upper 27% and a lower 27% (Kelley, 1939). This, however, assumes there is a large enough sample size. For the purpose of classroom music assessments, it is suggested to include all students by dividing the group into an equally divided upper 50% and lower 50%. If there is an uneven grouping of students, remove the middle-most student.
3. Calculate p_{i_high} (the item difficulty index for the high-ability group).
    a. Sum the total correct answers for the students in the high-ability group.
    b. Calculate the number of total students in the high-ability group.
    c. Divide the sum of the total correct answers for the students in the high-ability group by the total number of students in the high-ability group.
4. Calculate p_{i_low} (the item difficulty index for the low-ability group).
    a. Sum the total correct answers for the students in the low-ability group.
    b. Calculate the number of total students in the low-ability group.
    c. Divide the sum of the total correct answers for the students in the low-ability group by the total number of students in the low-ability group.
5. Calculate D_i (the item discrimination index).
    a. Subtract the item difficulty index of the low-ability group (p_{i_low}) from the item difficulty index of the high-ability group (p_{i_high}).

As an example, let us calculate the item discrimination index for Item 4. If we evaluate the ordered right/wrong matrix displayed in Table 8.6, we see that the students are evenly split (50/50) into a high-ability group and a low-ability group, consisting of nine students each. The grey shading indicates the high-ability group. In order to calculate the item difficulty index for the high-ability group (p_{i_high}), divide the sum of the total correct answers for the students in the high-ability group (nine) by the total number of students in the high-ability group (nine). The item difficulty index for the high-ability group is equal to 1.00. Substantively, 100% of the students in the high-ability group answered Item 4 correctly. In order to calculate the item difficulty index for the low-ability group (p_{i_low}), divide the sum of the total correct answers for the students in the low-ability group (three) by the total number of students in the low-ability group (nine). The item difficulty index for the low-ability group is equal to 0.33. Substantively, 33% of the students in the low-ability group answered Item 4 correctly. To calculate the item discrimination index for Item 4 (D_{i4}), subtract the item difficulty index of the low-ability group (p_{i_low} = *0.33*) from the item difficulty index of the high-ability group (p_{i_high} = 1.00). The item discrimination index for Item 4 (D_{i4}) is equal to 0.67.

The item discrimination indices for each of the 20 items are found in Table 8.6. In evaluating these indices, values range from −1.00 to 1.00. There are two pieces of information that can be gleaned from their values: (a) directionality and (b) range. Positively discriminating items (values greater than 0.00) indicate that the high-ability group more often answers the item correctly than the lower-ability group (high-ability group scores > lower-ability group scores). Negatively discriminating items (values less than 0.00) indicate that the low-ability group more often answers the item correctly than the high-ability group (high-ability group scores < low-ability group scores). Non-discriminating items (values equal or close to 0.00) indicate that there is no substantial difference between the high-ability group score and low-ability group score on the item (high-ability group scores = low-ability group scores). From a data analysis perspective, music teachers would hope to find their items to be positively discriminating, indicating that the students in the high-ability group are more often answering the questions correctly than the students in the low-ability group.

The range of the discrimination index quantifies the quality of the relationship between class ability and answering the item correctly. The values can be interpreted as follows:

- *Very good item*: 0.40–1.00 (use; class responses are trustworthy).
- *Reasonably good item*: 0.30–0.39 (consider revising item; consider investigating class responses).
- *Fairly good item*: 0.11–0.29 (revise item; definitely investigate class responses).
- *Poor item*: 0.00–0.10 (do not use item; class responses are not trustworthy).

Note that the relationship between the interpretative categories and number ranges for item discrimination is different than the interpretative categories and number ranges for item difficulty. In practice, item discrimination indices rarely exceed 0.50 due to various distributions in the relationships between item performance and total test scores.

Negatively discriminating items or items with low discrimination values provide evidence that, from a class perspective, something is wrong with either the testing context (e.g., mistake in the answer collection

Table 8.6 Calculation of item discrimination indices

		Agents																			
		I19	*I20*	*I18*	*I10*	*I11*	*I09*	*I14*	*I04*	*I06*	*I13*	*I16*	*I01*	*I05*	*I02*	*I03*	*I08*	*I12*	*I07*	*I17*	*I15*
Objects	**Student 4**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	Student 15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0
	Student 11	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0
	Student 17	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0
	Student 13	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	1
	Student 5	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0
	Student 10	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	1	0	0
	Student 9	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	0	0	0	0
	Student 6	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	1
	Student 18	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0
	Student 14	1	1	1	1	1	1	0	1	1	1	0	1	0	0	1	0	0	0	0	0
	Student 2	1	1	0	0	1	0	1	0	1	1	0	1	0	1	0	0	1	0	1	0
	Student 16	1	0	0	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	1	1
	Student 3	1	1	1	0	0	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0
	Student 8	1	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
	Student 12	1	1	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0
	Student 7	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0
	Student 1	0	0	1	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
	Item Sum Score	17	16	14	14	14	13	13	12	12	11	11	10	10	8	8	8	8	7	4	3

p_i	0.94	0.89	0.78	0.78	0.78	0.72	0.72	0.67	0.67	0.61	0.61	0.56	0.56	0.44	0.44	0.44	0.44	0.39	0.22	0.17
# correct high	9	9	9	9	8	9	8	9	9	5	8	7	7	6	6	7	4	5	2	2
p_{i_high}	1.00	1.00	1.00	1.00	0.89	1.00	0.89	1.00	1.00	0.56	0.89	0.78	0.78	0.67	0.67	0.78	0.44	0.56	0.22	0.22
# correct low	8	7	5	5	6	4	5	3	3	6	3	3	3	2	2	1	4	2	2	1
p_{i_low}	0.89	0.78	0.56	0.56	0.67	0.44	0.56	0.33	0.33	0.67	0.33	0.33	0.33	0.22	0.22	0.11	0.44	0.22	0.22	0.11
D_i	0.11	0.22	0.44	0.44	0.22	0.56	0.33	0.67	0.67	-0.11	0.56	0.44	0.44	0.44	0.44	0.67	0.00	0.33	0.00	0.11

Note. Grey shading indicates high-ability group.

process), the item (e.g., the writing style, misleading answers, wrong information), the instructional delivery underscoring the content of the item (e.g., content was never covered, did not convey answer), or the students' understanding of the content knowledge (e.g., class confusion about a topic, guessing). Therefore, these items should be removed from the testing context when considering the evaluation of student learning outcomes. Furthermore, engagement with the class about the testing context and class knowledge of the content is encouraged.

Person-Discrimination Indices

Person-discrimination indices are important for empirically exploring the quality of the response patterns of the individual students. Conceptually, person-discrimination can be thought of as a value that represents the frequency with which students responded correctly to items by varying groups of difficulty levels of items, such as comparing more difficult item response patterns to less difficult item response patterns, for example. Similar to item-discrimination, person-discrimination is represented by a *D*-value (discrimination value) and is calculated as follows:

$$D_p = p_{p\_easy} - p_{p\_difficult}$$

where D_p = discrimination of person *i*,
p_{p_easy} = person ability index of the less-difficult item group, and
$p_{p\_difficult}$ = person ability index of the more-difficult item group.

In order to arrive at the calculation of D_p, there are some considerations to be made and steps to go through:

1. Start by creating a right/wrong matrix, ensuring that the items are ordered from least difficult to most difficult.
2. Divide the items evenly into a more-difficult group and less-difficult group. Divide the groups into an equal 50/50 split (50% more-difficult and 50% less-difficult). If there is an uneven grouping of items, remove the middle-most item.
3. Calculate $p_{p\_less\_difficult}$ (the person ability index for the less-difficult item group).

    a. Sum the total correct answers for the items in the less-difficult item group.
    b. Calculate the number of total items in the less-difficult item group.
    c. Divide the sum of the total correct answers for the items in the less-difficult item group by the total number of items in the less-difficult item group.

4. Calculate $p_{p\_more\_difficult}$ (the person ability index for the more-difficult item group).
   a. Sum the total correct answers for the items in the more-difficult item group.
   b. Calculate the number of total items in the more-difficult item group.
   c. Divide the sum of the total correct answers for the items in the more-difficult item group by the total number of items in the more-difficult item group.
5. Calculate D_p (the person ability index).
   a. Subtract the person ability index of the more-difficult item group ($p_{p\_more\_difficult}$) from the person ability index of the less-difficult item group ($p_{p\_less\_difficult}$).

As an example, let us calculate the person ability index for Student 6. If we evaluate the ordered right/wrong matrix displayed in Table 8.7, we see that the items are evenly split (50/50) into a more-difficult item group and a less-difficult item group, consisting of ten items each. The grey shading indicates the less-difficult item group. In order to calculate the person ability index for the less-difficult item group ($p_{p\_less\_difficult}$), divide the sum of the total correct answers for the items in the less-difficult item group (nine) by the total number of items in the less-difficult item group (ten). The person ability index for the less-difficult item group is equal to 0.90. Substantively, Student 6 answered 90% of the items in the less-difficult item group correctly. In order to calculate the person ability index for the more-difficult item group ($p_{p\_more\_difficult}$), divide the sum of the total correct answers for the items in the more-difficult item group (three) by the total number of items in the more-difficult item group (ten). The person ability index for the more-difficult item group is equal to 0.30. Substantively, Student 6 answered 30% of the items in the difficult item group correctly. To calculate the person ability index for Student 6 (D_{s6}), subtract the person ability index of the more-difficult item group ($p_{p\_more\_difficult}$ = 0.30) from the person ability index of the less-difficult item group ($p_{p\_less\_difficult}$ = *0.90*). The item discrimination index for student 6 (D_{p6}) is equal to 0.60.

The person ability indices for each of the 18 students are found in Table 8.7.

Similar to the item discrimination indices, values range from –1.00 to 1.00. The positively discriminating persons (values greater than 0.00) indicate that the less-difficult item grouping was more often answered correctly than the more-difficult item grouping (more-difficult item group scores > less-difficult item group scores). Negatively discriminating persons (values less than 0.00) indicate that the more-difficult item grouping was more often answered correctly than the less-difficult item grouping (less-difficult item group scores < more-difficult item group scores).

Table 8.7 Calculation of person discrimination indices

		Agents																				*Person Sum Score*	p_p	*# correct less-dif*	p_{i_easy}	*# correct more-dif*	p_{i_dif}	D_p
		I19	*I20*	*I18*	*I10*	*I11*	*I09*	*I14*	*I04*	*I06*	*I13*	*I16*	*I01*	*I05*	*I02*	*I03*	*I08*	*I12*	*I07*	*I17*	*I15*							
Objects	**S4**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	**19**	0.90	10	1.00	9	0.90	0.10
	S15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	**18**	0.86	10	1.00	8	0.80	0.20
	S11	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	**16**	0.76	9	0.90	7	0.70	0.20
	S17	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0	**16**	0.76	9	0.90	7	0.70	0.20
	S13	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	1	**16**	0.76	9	0.90	7	0.70	0.20
	S5	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	**15**	0.71	9	0.90	6	0.60	0.30
	S10	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	1	0	0	**13**	0.62	10	1.00	3	0.30	0.70
	S9	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	0	0	0	0	**13**	0.62	9	0.90	4	0.40	0.50
	S6	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	1	**12**	0.57	9	0.90	3	0.30	0.60
	S18	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0	**12**	0.57	10	1.00	2	0.20	0.80
	S14	1	1	1	1	1	1	0	1	1	1	0	1	0	0	1	0	0	0	0	0	**11**	0.52	9	0.90	2	0.20	0.70
	S2	1	1	0	0	1	0	1	0	1	1	0	1	0	1	0	0	1	0	1	0	**10**	0.48	6	0.60	4	0.40	0.20
	S16	1	0	0	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	1	1	**9**	0.43	4	0.40	5	0.50	–0.10
	S3	1	1	1	0	0	1	1	0	0	1	0	0	0	0	0	0	1	1	0	0	**8**	0.38	6	0.60	2	0.20	0.40
	S8	1	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	**7**	0.33	6	0.60	1	0.10	0.50
	S12	1	1	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	0	**7**	0.33	3	0.30	4	0.40	–0.10
	S7	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	**6**	0.29	5	0.50	1	0.10	0.40
	S1	0	0	1	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	**5**	0.24	3	0.30	2	0.20	0.10

Non-discriminating items (values equal or close to 0.00) indicate that there is no substantial difference between the less-difficult item group scores and more-difficult item group scores by the student (less-difficult item scores = more-difficult item scores). From a data analysis perspective, music teachers would hope to find their students to be positively discriminating, indicating that the student is answering the less-difficult items correctly more frequently than the more-difficult items.

The range of the discrimination quantifies the quality of the relationship between item difficulty and individual student ability. The values can be interpreted as follows:

- *Good-quality responses*: 0.40–1.00 (use student outcome data; patterns in student responses are acceptable).
- *Reasonable-quality responses*: 0.30–0.39 (use student outcome data with some caution; consider investigating student responses; patterns in student responses are somewhat acceptable).
- *Fair-quality responses*: 0.11–0.29 (consider not using student outcome data; investigate student responses; patterns in student responses are somewhat unacceptable).
- *Poor-quality responses*: 0.00–0.10 (do not use student outcome data; definitely investigate student responses; student responses are unacceptable).

The values can be interpreted in the same manner as item discrimination indices. However, these indices are now student-centered (as opposed to class-centered) and provide a mechanism to investigate individual students and the adequacy of their learning. Negatively discriminating students or students with low discrimination values provide evidence that, from a student perspective, something is wrong with either the testing context (e.g., mistakes in the answer keying process, did not understand directions, skipped/did not answer items), the student himself or herself (e.g., did not engage the test, sick, distracted), the instructional delivery underscoring the content (e.g., missed instruction, did not study material covered), or the student's understanding of the content knowledge (e.g., individual student confusion about a topic, guessing). Therefore, these students should be removed from the testing context when considering the quality of the evaluation of student learning outcomes. Furthermore, engagement/intervention with the individual student or students' parents about the testing context and acquisition/understanding of instructional content is encouraged.

Distractor Analyses

In the event that the music educator finds item difficulty indices, person ability indices, item-discrimination indices, or person-discrimination indices that are problematic and warrant follow-up with either the class (for item-related abnormalities in the data) or individual students (for person-related abnormalities in the data), distractor analyses can be conducted as a means to further investigate more substantive issues pertaining to item and/or person response patterns.

Item Distractor Analyses

If we revisit the item difficulty (p_i) indices and item-discrimination (D_i) indices from Table 8.6, we can see that Item 13 is a suitable item for further investigation and possible class discussion. Item 13 demonstrated an item difficulty of 0.61 and an item-discrimination of –0.11. The item is considered average-difficulty; however, the discrimination index indicates that more lower-ability students answered the item correctly than high-ability students. If we revisit the original observations displayed in Table 8.1 and investigate the dichotomous responses of the high group and low group to Item 13 (Table 8.2), we can extrapolate the full details of the response patterns and display them in a table similar to the one reflected in Table 8.8. After evaluating the item distractor analysis for Item 13, we see that response option 3 and response option 4 were never selected. If we were to revise the item for future use, we may wish to eliminate both options 3 and 4 since all students clearly viewed them as inappropriate responses. The negative discrimination index indicates that more students from the low-ability group answered Item 13 correctly than students from the high-ability group. As we can see, 66.7% of the low-ability-group students answered the item correct compared to 55.6% of the students from the high-ability group. There is clearly something enticing in response option 2 for the students to respond to, as members of the low-ability group (33.3%) and high-ability group (44.4%) selected response option 2. Therefore, the teacher may wish to address why this occurred from a content perspective. If it is a problem with the item itself or the instructional delivery, the teacher should make adjustments to the item and/or instructional delivery for the next instructional/assessment cycle and then reevaluate the question after the completion of the next cycle's test. If it is a content-related concern, an open class discussion and instructional follow-up may be necessary to improve overall student understanding of the instructional content underscoring Item 13.

Person Distractor Analyses

If we revisit the person ability (p_p) indices and person-discrimination (D_p) indices from Table 8.7, we can see that Student 16 is a suitable person for further investigation and possible individual student conversation/intervention. Student 16 demonstrated a person ability of 0.43 and a person-discrimination of –0.10. Student 16 is considered an average-ability student; however, the discrimination index indicates that as member of the low-ability group, he or she answered more-difficult answers correctly and less-difficult items incorrectly than members of the high-ability group. If we revisit the original observations displayed in Table 8.1 and investigate

Table 8.8 Item distractor analysis for Item 13

	*Response option 1**		*Response option 2*		*Response option 3*		*Response option 4*	
	Frequency	*%*	*Frequency*	*%*	*Frequency*	*%*	*Frequency*	*%*
High Group	5	55.6%	4	44.4%	0	0.00%	0	0.00%
Low Group	6	66.7%	3	33.3%	0	0.00%	0	0.00%

Note. High-group students (in rank order) include 4, 15, 11, 17, 13, 5, 10, 9 and 6. Low-group students (in rank order) include Students 18, 14, 2, 16, 3, 8, 12, 7, and 1.

* indicates correct answer.

Table 8.9 Item distractor analysis for Item 13

Person Distractor Analysis for Student 16

			High Group		*Low Group*	
Item	p_i	*Observation*	*% Correct*	*% Incorrect*	*% Correct*	*% Incorrect*
Item 15	0.17	1	0.11	0.89	0.22	0.78
Item 17	0.22	1	0.22	0.78	0.22	0.78
Item 7	0.39	1	0.56	0.44	0.22	0.78
Item 2	0.44	1	0.67	0.33	0.22	0.78
Item 3	0.44	0	0.67	0.33	0.22	0.78
Item 8	0.44	0	0.67	0.33	0.22	0.78
Item 12	0.44	0	0.56	0.44	0.33	0.67
Item 1	0.56	0	0.78	0.22	0.33	0.67
Item 5	0.56	1	0.89	0.11	0.22	0.78
Item 13	0.61	1	0.56	0.44	0.67	0.33
Item 16	0.61	0	0.78	0.22	0.44	0.56
Item 4	0.67	0	1.00	0.00	0.33	0.67
Item 6	0.67	0	1.00	0.00	0.33	0.67
Item 9	0.72	0	1.00	0.00	0.44	0.56
Item 14	0.72	0	1.00	0.00	0.44	0.56
Item 10	0.78	1	1.00	0.00	0.56	0.44
Item 11	0.78	1	0.89	0.11	0.67	0.33
Item 18	0.78	0	1.00	0.00	0.56	0.44
Item 20	0.89	0	1.00	0.22	0.78	0.22
Item 19	0.94	1	1.00	0.00	0.89	0.11

the dichotomous responses of the less-difficult items and more- difficult items for Student 16 (Table 8.3), we can extrapolate the full details of the response patterns and display them in a table similar to the one reflected in Table 8.9. After evaluating the person distractor analysis for Student 16, we see the student answered the four most difficult items correctly (Items 15, 17, 7 and 2), which is completely unexpected. We also see that there is an interesting pattern of correct responses and incorrect responses as related to the ordering of items from more-difficult to less-difficult. In the context of an authentic assessment scenario, there may be any number of student-centered reasons for the results, such as implications for understanding of the content underlying the items, opportunity-to-learn considerations, attendance, or a unique interpretation of instructional content delivery. Table 8.9 provides new empirical insights into how this individual student demonstrates learning of the instructional material, and the results of the person-discrimination indices have brought needed attention to this student for possible intervention. In this instance, a one-on-one meeting with the student is encouraged in order to better understand their unique interpretation of the content and/or testing considerations that may have affected their performance on the test.

Summarizing Results

As Wright and Stone (1979) note, there are three important properties of any test (T): (a) test height (H), (b) test width (W), and (c) test length (L). Test height (H) refers to the average difficulty of the test items. Test width (W) refers to the ability range of the persons taking the test. Test length (L) refers to the number of items used in the test. According to Wright, this information can be extracted and reported as follows: T (H, W, L).

In order to report the test height (H), we can calculate the average difficulty of the test items. If we calculate the average item difficulty (p_i) values across the p_i row in Table 8.6, we get a value of 0.59. Using the same item difficulty interpretations listed above, we can conclude that the test is of average difficulty. In order to report the test width (W), we can evaluate the range of person ability. According to Table 8.7, the range of person ability is from 0.24 to 0.90. Using the same person ability value interpretations listed above, we can conclude that the persons range from low-ability to high-ability, with an average of 0.56 (average-ability). More specifically, persons consisted of 1 low-ability student (Student 1), 12 average-ability students (Students 7, 12, 8, 3, 16, 2, 14, 18, 6, 9, 10, and 5), and 5 high-ability students (Students 13, 17, 11, 15, and 4). Using Wright's suggestion, the test-centered properties can be reported as:

Instrument Timbre Test (0.59, 0.24–0.90, 20).

This report includes all items and persons. If we exclude items with low or negative discrimination indices (Items 13, 12, 17; see Table 8.6) and persons with low or negative discrimination indices (Persons 4, 16, 12, 1), the test-centered properties can be reported as:

Instrument Timbre Test (0.62, 0.29–0.86, 17).

Validity, Reliability, and Fairness Follow-Up Considerations

We noted that there were some quality concerns with the patterns of responses related to some items and some persons. In these cases, the items and persons with unexpected response patterns are not necessarily a quality representation of the overall assessment context. Therefore, from a student-centered perspective, it is the music educator's ethical responsibility to consider whether there are any validity, reliability, or fairness concerns that are associated with these unexpected outcomes. It is suggested that the questions aligned to each of the validity, reliability, and fairness quality indicators described in Chapter 5 be revisited and considered as possible influencers for the outcomes of the unexpected response patterns observed in the testing outcomes.

A Note on Polytomous Items

The example outlined in this chapter was a multiple-choice test that resulted in dichotomous (correct/incorrect) responses. In many instances, particularly in the context of music performance assessments, rating scales or rubrics may be used to evaluate student performance. For these types of evaluation instruments, there may be more than two response categories associated with the criteria. As an example, a Likert-type rating scale may include four response categories such as *Strongly Disagree, Disagree, Agree*, and *Strongly Agree*. A rubric may include four response categories such as *Emerging, Approaching Standard, Meeting Standard*, and *Exceeding Standard*. In instances where response categories include more than two options, they are said to be **polytomous**. For polytomous items, item difficulty, person ability, item-discrimination, and person-discrimination indices are calculated in the same manner as dichotomous items (as described in this chapter). Fundamentally, each of these indices are calculated using proportions. For example:

Item difficulty is calculated as:

$$p_i = \frac{R_i}{T_i},$$

Person ability is calculated as:

$$p_p = \frac{R_p}{T_p},$$

Item-discrimination is calculated as:

$$D_i = \left(\frac{sum correct\ answers\ high - ability\ group}{total\ students\ high - ability\ group} \right) - \left(\frac{sum correct\ answers\ low - ability\ group}{total\ students\ low - ability\ group} \right)$$

Person-discrimination is calculated as:

$$D_p = \left(\frac{sumcorrect\ answers\ less\ difficult\ item\ group}{total\ item\ less\ difficult\ group} \right) - \left(\frac{sumcorrect\ answers\ more\ difficult\ item\ group}{total\ item\ more\ difficult\ group} \right)$$

As a result, the calculations will hold true with more than two response options. However, there are three important requirements for this to work. First, all of the criteria must have the same response category structure. As an example, having one criterion with a response category structure of *Approaching Standard/Meeting Standard* and another criterion with a response category structure of *Approaching Standard/Meeting Standard/Exceeding Standard* will not work. Second, the response categories must be coded in the same, ascending order. As an example, if there are four categories for every criterion, the lowest category should be labeled 1, the second category should be labeled 2, the third category should be labeled 3, and the highest category should be labeled 4. Lastly, the same label should represent each response category across all criteria. As an example, if each criterion has the response category *Strongly Disagree, Disagree, Agree*, and *Strongly Agree, Strongly Disagree* will always be coded as a 1 for every criterion, *Disagree* will always be coded as a 2 for every criterion, *Agree* will always be coded as a 3 for every criterion, and *Strongly Agree* will always be coded as a 4 for every criterion. If there are four response categories, the total possible amount a student can score is 4 points.

Summary

It is important to clearly communicate to administrators and stakeholders the teaching and learning occurring in the music classroom with empirical data. Additionally, it is important to examine assessment data using empirical methods as a means to explore the quality of students' learning. There are four important indices that provide insights into the outcomes and quality of testing data: (a) item difficulty indices, (b) person ability indices, (c) item-discrimination indices, and (d) person-discrimination indices. Item difficulty is an important index for exploring the proportion of students who answered an item correctly and incorrectly. Person ability is an important index for exploring the proportion of items that were answered correctly or incorrectly by an individual student. Item-discrimination is an important index for empirically exploring the quality of the response patterns of the items. Person-discrimination is an important index for empirically exploring the quality of the response patterns of the individual students. In the case that there are abnormalities in the outcome data, item and person distractor analyses as well as the qualitative considerations for the validity, reliability, and fairness of student outcomes are important tools to inform future teaching and learning processes and improving future test uses from both a class-centered perspective and an individual student-centered perspective.

Activity Worksheet

	Item 01	*Item 02*	*Item 03*	*Item 04*	*Item 05*	*Item 06*	*Item 07*	*Item 08*	*Item 09*	*Item 10*
Student 01	1	1	0	1	1	0	1	1	1	1
Student 02	1	0	1	1	1	0	1	1	0	1
Student 03	1	0	1	1	1	0	1	1	0	0
Student 04	1	0	0	0	1	1	1	1	1	0
Student 05	1	1	0	1	0	1	0	1	0	0
Student 06	1	0	0	1	1	0	1	1	1	0
Student 07	1	1	1	0	1	0	1	0	0	0
Student 08	1	1	1	0	1	0	0	0	0	0
Student 09	1	1	0	0	0	0	0	1	0	0
Student 10	0	1	1	0	0	0	0	0	1	0

Above is a sample right/wrong matrix. A total of 100 observations were collected. The data are dichotomously scored and represent 10 students responding to 10 items. Complete the following:

1. Sum the person scores.
2. Sum the item scores.
3. Rearrange the right/wrong matrix by ordering the items from least difficult to most difficult and persons from highest-achieving to lowest-achieving.
4. Draw a diagonal line from the top right part of the matrix down to the bottom left part of the matrix.
5. Document any items and persons where unexpected responses are occurring.
6. Calculate an item difficulty index for each item.
7. Which items are considered easy? Which items are considered average-difficulty? Which items are considered difficult?
8. Calculate a person ability index for each person.
9. Which persons are considered high-ability? Which persons are considered average-ability? Which persons are considered low-ability?
10. Split the items into a 50% less-difficult group and 50% more-difficult group.
11. Calculate an item-discrimination index for each item.

12. Which items are considered very good items? Which items are considered reasonably good items? Which items are considered fairly good items? Which items are considered poor items?
13. Calculate a person-discrimination index for each person.
14. Which persons are considered to have provided good-quality responses? Which persons are considered to have provided reasonable-quality responses? Which persons are considered to have provided fair-quality responses? Which persons are considered to have provided poor-quality responses?
15. Choose one negatively discriminating item/poor item and construct an item distractor analysis table.
16. Choose one negatively discriminating person/poor-quality-response person and construct a person distractor analysis table.
17. Report the test height, test width, and test length including all items and persons.
18. Report the test height, test width, and test length after excluding the negatively discriminating item/poor items and the negatively discriminating person/poor-quality-response persons.

Notes

1. When evaluating testing data, the term *person* is used to reflect the group of students, or objects of measurement. When broadly discussing the syntax of the data, the term *person* will be used. When discussing the individual test performance or student in context of the classroom, the term *student* will be used.
2. When providing an interpretation of item *difficulty* or person *ability*, it is from a strict data analysis perspective, not necessary the teacher's interpretation of "difficulty" of an item or "ability" of a student in the context of teaching and learning/classroom expectations.

References

Hart, K. (2003). From an administrator's perspective: Practical survival skills for music educators. *Music Educators Journal*, *90*(2), 41–45.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, *30*, 17–24.

Lord, F. M. (1952). The relationship of the reliability of multiple-choice test to the distribution of item difficulties. *Psychometrika*, *18*, 181–194.

National Comprehensive Center for Teacher Quality. (2011). *Measuring teachers' contributions to student learning growth and non-tested grades and subjects*. Washington, DC: Educational Testing Service.

Wesolowski, B. C. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal*, *101*(1), 77–85.

Wesolowski, B. C. (2015). Tracking student achievement in music performance: Developing student learning objectives for growth model assessments. *Music Educators Journal*, *102*(1), 39–47.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.