SECOND LANGUAGE TESTING FOR STUDENT EVALUATION AND CLASSROOM RESEARCH

Student Workbook

Greta Gorsuch and Dale T. Griffee

Second Language Testing for Student Evaluation and Classroom Research

Student Workbook

Second Language Testing for Student Evaluation and Classroom Research

Student Workbook

by

Greta Gorsuch

and

Dale Griffee

INFORMATION AGE PUBLISHING, INC.
Charlotte, NC • www.infoagepub.com

Library of Congress Cataloging-in-Publication Data

CIP record for this book is available from the Library of Congress
http://www.loc.gov

ISBNs: 978-1-64113-017-2 (Paperback)

978-1-64113-018-9 (ebook)

Printed in the United States of America

STUDENT WORKBOOK CONTENTS

STUDENT WORKBOOK
CHAPTER 1

NORM-REFERENCED TESTS

TEST YOURSELF

Working alone or with a study partner, ask and answer these questions:

1. What does TREE stand for?
2. What is a norming group?
3. What is Item Facility (IF)?
4. What is Item Discrimination (ID)?
5. What are five advantages of NRTs?
6. What are five disadvantages of NRTs?
7. What does the term SEM mean?
8. Explain what local validation means.
9. What is a measurement model?
10. What is domain theory? Can you think of an example?

Second Language Testing for Student Evaluation and Classroom Research Student Workbook, pp. 1–9

DISCUSSION QUESTIONS

1. Under what conditions would you be happy that your students took a norm-referenced test (NRT) and that you had access to their scores?

2. Under what circumstances might you make an NRT?

3. Your school director wants to evaluate your language program. In a meeting, he suggests using a standardized test he can buy to measure students' proficiency when they enter the program and again when they leave. He believes these test scores will help students find a job. How do you respond?

4. You are working in a small language school. For some time, teachers have been complaining that their classes have mixed levels of students in them and that is making it hard for them to teach. The school director asks you for your advice to deal with this problem because she knows you have taken a testing course. She wishes to know whether you could make some kind of test that could be given to all new students to determine the level of the class to which they should be assigned. What are some possible recommendations you can make?

APPLICATION TASKS

1. An Item Facility (IF) is the percentage of items answered correctly. Calculate IFs for a simulated test with six students and seven items using the data in Table 1.1. The students (sometimes called test candidates) are in the first column. The next seven rows are scores for the students on each NRT item. If there is a 1, it means the student got the item right. If there is a 0, it means the student got the item wrong. Thus, for student 1, we can see that on item 1, she got the item right, but that on item 2 she got it wrong. This arrangement allows us to calculate and display IFs for each item. The formula is $IF = N_{correct} / N_{total}$. For item 1, the IF is .67 (4 divided by 6). In other words, 67% of the students got item 1 correct. Now, calculate the IFs for items 2 through 10.

2. Which items in Table 1.1 are easy? Which are in the middle? Which are difficult? Given the general guideline that good NRT items

have IFs between .2 and .8, which items would you keep? Why? Which items would you revise? Why?

3. Table 1.2 shows a different group of students who took an NRT with 100 total items. Using Table 1.2, calculate IFs and IDs for items 2 through 8. Note that the new group of students is now arranged in three levels by total score to identify IF upper and IF lower. For example, this means that Al (not his real name) got a total score of 85, calculated from adding up all of the correct answers from items 1 through 100 on. The lowest scoring student on the whole test was Ted. The formula is ID = $IF_{upper} - IF_{lower}$. Item 1 has been done for you. Note that 75% of the highest group answer item 1 correctly. Thus, the IF for the upper group is .75. Only 25% of the lowest group answered item 1 correctly, with an IF of .25. (75 minus .25 = .5, the ID for the item).

Table 1.1.
NRT Item-Level Data Arranged to Calculate Item Facility

Items	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10, etc.*
Student 1	1	0	1	1	1	1	1	0	1	0
Student 2	1	1	0	1	1	1	1	0	1	0
Student 3	1	0	1	0	1	1	0	0	1	1
Student 4	0	1	1	1	1	1	1	0	1	0
Student 5	1	0	0	1	1	1	0	0	1	0
Student 6	0	1	1	0	0	0	0	1	1	0
N correct	4									
N total	6									
IF =	.67									

Table 1.2.
NRT Items Arranged to Calculate Item Facility and Item Discrimination

Items	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8, etc.*	*Total Score*
Al	1	1	1	0	1	1	0	1	85
Wanda	1	1	1	0	0	1	0	0	82
Emily	1	1	0	0	1	1	0	0	81
Peter	0	1	1	0	1	1	0	1	80
Anna	1	1	1	0	1	0	0	1	75
Tom W.	0	1	0	1	0	1	0	1	74
Tom Z.	0	1	0	1	1	0	0	0	72
Neil	0	0	1	0	1	1	0	1	70
Joan	1	0	0	1	0	0	0	0	66
Shirley	0	1	0	0	1	0	0	0	65
Georgia	1	1	1	0	0	0	0	0	55
Jean	0	1	0	1	0	0	0	0	52
Kenny	0	1	0	0	0	0	1	0	50
Ted	0	1	0	1	0	0	1	0	48
IF_{upper}	.75								
IF_{lower}	.25								
ID	.50								

5. Which test items in Table 1.2 would you keep as good items? Why? Which items would you look at and revise? Why? How might you revise an item?
6. Following the format in Table 1.2, find item-level and total scores from an NRT test and calculate the IFs and IDs.

TEST YOURSELF ANSWERS

1. What does TREE stand for?
   Teacher/Researcher/Educator/Evaluator

2. What is a norming group?
   It's the group on which a norm-referenced (NRT) test is created.

3. What is Item Facility (IF)?
   An IF is the percentage of students who answer a test item correctly.

4. What is Item Discrimination (ID)?
   Item Discrimination shows the extent to which a test item separates high-scoring students from low-scoring students.

5. What are five advantages of NRTs?

   - They can measure language proficiency.
   - They level the playing field, meaning that the test is the same for all students.
   - Scores from an NRT can be used to characterize or describe students from a class.
   - We can buy an NRT from a test company.
   - A commercially produced NRT can include supplemental material such as practice test forms, CDs, and a test manual discussing the development of the test.

6. What are five disadvantages of NRTs?
   - They are a poor measure of learning in a particular program.
   - They require added test security.
   - Their scores have to be interpreted.
   - They measure traits, and traits are static and not dynamic.
   - They encourage belief that our knowledge is inborn rather than worked for.

7. What is an SEM?
   SEM stands for standard error of measurement and is a band around each person's NRT test score showing how high or low the score might actually be.

8. Explain what local validation means.
   Gathering evidence that the test measures your students accurately and appropriately.

9. What is a measurement model?
   A theory of underlying assumptions of a test.

10. What is domain theory?
    The theory of what the test (any test, not just an NRT) is purporting to measure. Domain theory examples include: (A) listening comprehension requires both top-down and bottom-up processing, (b) guessing unknown words from context is an important skill for independent reading, (c) learners writing in a foreign language need to learn to

write for different audiences, (d) communicative competence in speaking means knowing what to say depending on whom you are talking to and for what reason, and (e) the list is endless!

DISCUSSION QUESTIONS ANSWERS

1. Under what conditions would you be happy that your students took a norm-referenced test (NRT) and that you had access to their scores?
   When you want to make admissions or placement decisions, scores from an NRT is useful. Another use is comparing your students to other students in your program, past and present, who have also taken the test. This information may be used to project long-term responses you think your program should make in response to changes in the student population attending your school. Finally, if you want to determine that two or more groups of students are generally the same in language ability at the beginning of a research project, norm-reference test scores are useful.

2. Under what circumstances might you make an NRT?
   Some schools wish to make an NRT for admissions or placement to save money. There is also an argument that a placement test should have a stronger relationship to the program outcomes than commercially available NRTs generally do.

3. Your school director wants to evaluate your language program. In a meeting, he suggests using a standardized test he can buy to measure students' proficiency when they enter the program and again when they leave. He believes these test scores will help students find a job. How do you respond?
   The director has a point that graduates with high scores on a well-known NRT make a positive impression on prospective employers. However, an NRT should not be used to measure learning over time. Most NRTs will not detect improvement on well-defined program outcomes. Suggest to the director that a well-designed CRT be used to show learning over time, and report improvement in terms of the outcomes. Students can take the well-known NRT as an option to improve their job search portfolios.

4. You are working in a small language school. For some time, teachers have been complaining that their classes have mixed levels of students in them, and that is making it hard for them to teach. The school director asks you for your advice to deal with this problem

because she knows you have taken a testing course. She wishes to know whether you could make some kind of test that could be given to all new students to determine the level of the class to which they should be assigned. What are some possible recommendations you can make?

It is possible to buy or make an NRT to give to students when they start in a school. The scores can be used to make placement decisions. You may wish to suggest that you be paid extra or given teaching release to get consensus on the program outcomes, and then to find or make a placement NRT that has some relationship to the program outcomes.

APPLICATION TASKS ANSWERS

1. An IF (Item Facility) is the percentage of items answered correctly. Calculate IFs for a simulated test with six students and seven items using the data in Table 1.3. The students (sometimes called test candidates) are in the first column. The next seven rows are scores for the students on each NRT item. If there is a 1, it means the student got the item right. If there is a 0, it means the student got the item wrong. Thus, for student 1, we can see that on item 1, she got the item right, but that on item 2 she got it wrong. This arrangement allows us to calculate and display IFs for each item. The formula is IF = $N_{correct}$ / N_{total}. For item 1, the IF is .67 (4 divided by 6). In other words, 67% of the students got item 1 correct. Now, calculate the IFs for items 2 through 10.

Table 1.3.
NRT Item-Level Data Arranged to Calculate Item Facility Answers

Items	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10, etc.*
Student 1	1	0	1	1	1	1	1	0	1	0
Student 2	1	1	0	1	1	1	1	0	1	0
Student 3	1	0	1	0	1	1	0	0	1	1
Student 4	0	1	1	1	1	1	1	0	1	0
Student 5	1	0	0	1	1	1	0	0	1	0
Student 6	0	1	1	0	0	0	0	1	1	0
N correct	4	3	4	4	5	5	3	1	6	1
N total	6	6	6	6	6	6	6	6	6	6
IF =	.67	.50	.67	.67	.83	.83	.50	.17	1.00	.17

2. Which items in Table 1.3 are easy? Which are in the middle? Which are difficult items? Given the general guideline that good NRT items have IFs between .2 and .8, which items would you keep? Why? Which items would you revise? Why?
   Items 5, 6, and 9 are easy for the students. Items 1, 2, 3, 4, and 7 are in the middle. Items 8 and 10 are difficult for students. Items 1, 2, 3, 4, and 7 are within the suggested range and should be retained. Items 5, 6, and 9 are too easy, whereas items 8 and 10 are too difficult. Thus, the items may not function to spread students out on a score continuum. These five items should be revised.
3. Table 1.4 shows a different group of students who took an NRT with 100 total items. Using Table 1.4, calculate IFs and IDs for items 2 through 8. Note that the new group of students is now arranged in three levels by total score to identify IF upper and IF lower. For example, this means that Al (not his real name) got a total score of 85, calculated from adding up all of the correct answers from items 1 through 100 on. The lowest scoring student on the whole test was Ted. The formula is ID = $IF_{upper} - IF_{lower}$. Item 1 has been done for you. Note that 75% of the highest group answer item 1 correctly. Thus, the IF for the upper group is .75. Only 25% of the lowest group answered item 1 correctly, with an IF of .25 (.75 – .25 = .5, the ID for the item).

Table 1.4.
NRT Items Arranged to Calculate Item Facility and Item Discrimination Answers

Items	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8, etc.*	*Total Score*
Al	1	1	1	0	1	1	0	1	85
Wanda	1	1	1	0	0	1	0	0	82
Emily	1	1	0	0	1	1	0	0	81
Peter	0	1	1	0	1	1	0	1	80
Anna	1	1	1	0	1	0	0	1	75
Tom W.	0	1	0	1	0	1	0	1	74
Tom Z.	0	1	0	1	1	0	0	0	72
Neil	0	0	1	0	1	1	0	1	70
Joan	1	0	0	1	0	0	0	0	66

(Table continues on next page)

Table 1.4.
(Continued)

Items	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8, etc.*	*Total Score*
Shirley	0	1	0	0	1	0	0	0	65
Georgia	1	1	1	0	0	0	0	0	55
Jean	0	1	0	1	0	0	0	0	52
Kenny	0	1	0	0	0	0	1	0	50
Ted	0	1	0	1	0	0	1	0	48
IF_{upper}	.75	1.00	.75	.00	.75	1.00	.00	.50	
IF_{lower}	.25	1.00	.25	.50	.00	.00	.50	.00	
ID	.50	.00	.50	–.50	.75	1.00	–.50	.50	

4. Which test items in Table 1.4 would you keep as good items? Why? Which items would you look at and revise? Why? How might you revise an item?
   Items with IDs above .20 are 1, 3, 5, 6, and 8. These items are effective to discriminate between high- and low-scoring students, as measured by their total test scores. Items 2, 4, and 7 do not discriminate between high- and low-scoring students. Item 2 appears to be easy for both high- and low-scoring students, and so the item might be made more difficult for low- scoring students. Items 4 and 7 need to have their wording checked. Something about the wording or the test directions is throwing the high-scoring students off and somehow benefitting the low-scoring students.
5. Following the format in Table 1.4, find item-level and total scores from an NRT test and calculate the IFs and IDs.