Validity Argument in Language Testing

Language tests play pivotal roles in education, research on learning, and gate-keeping decisions. The central concern for language testing professionals is how to investigate whether or not tests are appropriate for their intended purposes. This book introduces an argument-based validity framework to help with the design of research that investigates the validity of language test interpretation and use. The book presents the principal concepts and technical terms and then shows how they can be implemented in practice through a variety of validation studies. It also demonstrates how argument-based validity intersects with technology in language testing research and highlights the use of validity argument for identifying research questions and interpreting the results of validation research. Use of the framework helps researchers in language testing to communicate clearly and consistently about technical issues with each other and with researchers of other types of tests.

Carol A. Chapelle is Distinguished Professor in Applied Linguistics at Iowa State University. She received the 2012 Cambridge-International Language Testing Association Lifetime Achievement Award and the 2020 Test Validity Research and Evaluation SIG Senior Scholar Award from the American Educational Research Association. She is Co-editor of the Cambridge Applied Linguistics Series and is Founding Editor of the Encyclopedia of Applied Linguistics.

Erik Voss is Lecturer in Applied Linguistics & TESOL at Teachers College, Columbia University. He is a member of the International Language Testing Association’s (ILTA) executive board and has co-authored the book Flipping Academic English Language Learning: Experiences from an American University (2019).
Research Genres
Exploration and Applications
John M. Swales
Critical Pedagogies and Language Learning
Edited by Bonny Norton and Kelleen Toohey
Exploring the Dynamics of Second Language Writing
Edited by Barbara Kroll
Understanding Expertise in Teaching
Case Studies of Second Language Teachers
Amy B. M. Tsui
Criterion-Referenced Language Testing
James Dean Brown and Thom Hudson
Corpora in Applied Linguistics
Susan Hunston
Pragmatics in Language Teaching
Edited by Kenneth R. Rose and Gabriele Kasper
Cognition and Second Language Instruction
Edited by Peter Robinson
Research Perspectives on English for Academic Purposes
Edited by John Flowerdew and Matthew Peacock
Computer Applications in Second Language Acquisition
Foundations for Teaching, Testing and Research
Carol A. Chapelle
Validity Argument in Language Testing
Case Studies of Validation Research

Edited by

Carol A. Chapelle
Iowa State University

Erik Voss
Teachers College, Columbia University
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Contributors</td>
<td>xiv</td>
</tr>
<tr>
<td>Series Editor’s Preface</td>
<td>xv</td>
</tr>
<tr>
<td>1 Introduction to Validity Argument in Language Testing and Assessment</td>
<td>1</td>
</tr>
<tr>
<td>Carol A. Chapelle and Erik Voss</td>
<td></td>
</tr>
<tr>
<td><strong>PART I BASIC CONCEPTS AND USES OF VALIDITY</strong></td>
<td></td>
</tr>
<tr>
<td><strong>ARGUMENT IN LANGUAGE TESTING AND ASSESSMENT</strong></td>
<td>17</td>
</tr>
<tr>
<td>2 Understanding Argument-Based Validity in Language Testing</td>
<td>19</td>
</tr>
<tr>
<td>Carol A. Chapelle and Hye-won Lee</td>
<td></td>
</tr>
<tr>
<td>Ahmet Dursun and Zhi Li</td>
<td></td>
</tr>
<tr>
<td><strong>PART II INVESTIGATING SCORE INTERPRETATIONS</strong></td>
<td>71</td>
</tr>
<tr>
<td>4 Domain Definition Inference for a Virtual Interactive Aviation English Test (VIAET) for Military Air Traffic Controllers</td>
<td>73</td>
</tr>
<tr>
<td>Moonyoung Park</td>
<td></td>
</tr>
<tr>
<td>5 Support for the Evaluation Inference: Investigating Conditions for Rating Responses on a Test of Academic Oral Language</td>
<td>96</td>
</tr>
<tr>
<td>Hyejin Yang</td>
<td></td>
</tr>
<tr>
<td>6 Generalization Inference for a Computer-Mediated Graphic-Prompt Writing Test for ESL Placement</td>
<td>120</td>
</tr>
<tr>
<td>YunDeok Choi</td>
<td></td>
</tr>
</tbody>
</table>
Contents

7 The Telephone Standard Speaking Test: An Outside Evaluator’s Investigation of a Rebuttal to the Generalization Inference 154
   Rie Koizumi

8 The Explanation Inference for a Test of Academic English Collocational Ability 176
   Erik Voss

9 Extrapolation of the Meaning of Grades on Writing Tasks in an ESL Writing Class to Success on Tasks in University Courses 204
   Jooyoung Lee

PART III INVESTIGATING SCORE USES AND CONSEQUENCES 233

10 Justifying the Interpretation and Use of an ESL Writing Final Examination 235
    Heesung Jun

11 An Argument for Use of a Test of Productive Grammatical Ability for Placement into Writing Classes 264
    Yoo-Ree Chung

12 Investigating the Consequences of an ESL Placement Test 294
    Zhi Li

PART IV CONCLUSION 323

13 Validation Research and Its Limits 325
    Carol A. Chapelle and Erik Voss

Index 345
Figures

2.1 Diagram showing one part of a Toulmin (2003) argument diagram (adapted from Toulmin, 2003, p. 94) page 24
2.2 Diagram showing the structure of the I/UA for the test of collocational ability presented in Chapter 8 37
3.1 Distribution of publication years 51
3.2 Distribution of affiliations of corresponding authors on the seventy studies in language testing using argument-based validity 52
3.3 Distribution of tests/assessment types investigated on the seventy studies in language testing using argument-based validity 53
3.4 Distribution of the 168 individual inferences from the 70 studies in language testing using argument-based validity 54
4.1 An illustration of the grounds, claims, and inferences in the interpretation/use argument for the prototype VIAET test 77
4.2 The four-stage ECD process used for development of the VIAET 79
4.3 Virtual simulation of a military air base in the VIAET created using Second Life 83
5.1 Interpretation/use argument structure for OECT in support of R-PLAT 99
5.2 An explanatory sequential design of the current study 103
5.3 The OPI rating page in R-PLAT (Yang, 2016, p. 55) 105
5.4 Diagnostic descriptors for fluency (Yang, 2016, p. 56) 106
6.1 Diagram summarizing the interpretation/use argument for scores on the graphic-prompt writing tasks with the inference, warrant, assumptions, and backing covered in this chapter 130
6.2 Configuration of Task 1 (Essay 1) 136
7.1 Structure of the interpretation and use argument for this chapter 158
7.2 Structure of the generalization inference in the interpretation and use argument for this chapter 159
7.3 Frequency of the first and second test form scores 166
List of Figures

7.4 Scatterplot of the scores in the first test form (X-axis) and in the second test form (Y-axis) 168
8.1 Theorized relationships of collocational ability with vocabulary size and reading comprehension in a nomological network 179
8.2 Interpretation/use argument structure for the English collocational ability test 180
8.3 Mixed-methods embedded and sequential explanatory research design for the explanation inference 182
8.4 Directions and items on the collocational ability test 185
8.5 Vocabulary size test sample questions 187
8.6 Boxplot showing score distributions for three proficiency levels on the collocational ability test 192
8.7 Scatterplot showing the relationship between rank order of item facility for collocation frequency (high = most frequent) and rank order of item facility (high = easiest) 193
8.8 Responses about thinking in academic English while taking the test 195
8.9 Responses comparing language on the test and in university textbooks 195
9.1 Claims and inferences in the interpretation/use argument 206
10.1 Interpretation/use argument structure for the web-researching-to-write test 237
10.2 Visual model of convergent mixed-methods design procedures 245
10.3 Screenshot of the web-researching-to-write test 246
11.1 Interpretation/use argument structure 266
11.2 Embedded mixed-methods design for the validation of score use and consequences of the academic grammar test 273
12.1 An interpretation/use argument for the English placement test 296
12.2 Mixed-methods sequential explanatory design 298
12.3 A simplified fully recursive model of the relationship among the EPT latent factor scores, motivational factors, and cumulative GPA, with standardized parameters 313
Tables

2.1 Important concepts for investigating the validity of language test interpretation and use ........................................ page 22

2.2 Functions of argument-based validity in language testing .................................................................................... 32

2.3 Contrasts between the arguments developed by Bachman and Palmer (2010) and by Kane as used in language assessment research ........................................................................................................ 34

3.1 Definitions and examples of all coded features ............................................................................................... 48

3.2 Collection of language testing research using an argument-based approach to validation ........................................... 52

3.3 Descriptive statistics of the number of inferences studied in three types of publications ........................................... 54

3.4 Summary of inferences being studied and methodologies used to investigate warrants and assumptions with each inference ........................................................................................................ 59

4.1 An example of task shell for a flight plan listening task .................................................................................... 88

4.2 Final blueprint for the Virtual Interactive Tasks for Aviation English assessment ................................................. 89

4.3 Interpretation of results for supporting assumptions underlying the domain definition inference ......................... 91

5.1 The averages of the experienced and new rater groups’ responses to the clarity of R-PLAT ........................................ 109

5.2 The average of both rater groups’ comfort level with R-PLAT ........................................................................ 111

5.3 The average level of both rater groups’ responses to the effectiveness of R-PLAT .................................................... 112

5.4 The average of both rater groups’ satisfaction with R-PLAT ................................................................................ 113

5.5 Summary of assumptions, backing, and degree of support for the evaluation inference ...................................... 115

6.1 Summaries of the selected previous studies on graphic-prompt writing tasks .......................................................... 122

6.2 Gender and L1 backgrounds of the student participants by group ........................................................................ 135

6.3 Descriptive statistics for composite and analytic scores on Tasks 1 and 2 ............................................................ 141

6.4 G-study estimated variance components for the composite and four analytic criteria scores ...................................... 142
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5</td>
<td>Estimated absolute variances and dependability indices for the composite and the analytic scores from the two-task and three-rater design</td>
<td>143</td>
</tr>
<tr>
<td>6.6</td>
<td>Summary of assumptions, backing, and degree of support for the generalization inference</td>
<td>146</td>
</tr>
<tr>
<td>7.1</td>
<td>Descriptive statistics of the first and second test forms</td>
<td>165</td>
</tr>
<tr>
<td>7.2</td>
<td>Percentages of test takers who received a level in the current study (N = 55) and average TSST taking population data (ALC, 2014)</td>
<td>166</td>
</tr>
<tr>
<td>7.3</td>
<td>Cross tabulation of two scores on the first and second test forms of the same participants: number of participants who belonged to each level</td>
<td>169</td>
</tr>
<tr>
<td>7.4</td>
<td>Percentages of test takers who selected each option</td>
<td>170</td>
</tr>
<tr>
<td>7.5</td>
<td>Summary of assumptions, backing, and degree of support for the generalization inference</td>
<td>171</td>
</tr>
<tr>
<td>7.6</td>
<td>Participants in the study by group</td>
<td>183</td>
</tr>
<tr>
<td>7.7</td>
<td>Summary of scoring methods</td>
<td>186</td>
</tr>
<tr>
<td>7.8</td>
<td>Descriptive statistics for the collocational ability test for all levels</td>
<td>191</td>
</tr>
<tr>
<td>7.9</td>
<td>Descriptive statistics for the reading test for all levels</td>
<td>193</td>
</tr>
<tr>
<td>7.10</td>
<td>Descriptive statistics for the vocabulary test for all levels</td>
<td>193</td>
</tr>
<tr>
<td>7.11</td>
<td>Summary of assumptions, backing, and degree of support for the explanation inference</td>
<td>197</td>
</tr>
<tr>
<td>7.12</td>
<td>Summary of warrants, assumptions, and backing associated with each inference in the interpretive argument for the English A writing assessment</td>
<td>208</td>
</tr>
<tr>
<td>7.13</td>
<td>Participants and their type of participation</td>
<td>213</td>
</tr>
<tr>
<td>7.14</td>
<td>Assignment sheets collected from English A</td>
<td>215</td>
</tr>
<tr>
<td>7.15</td>
<td>Assignments collected from content courses</td>
<td>216</td>
</tr>
<tr>
<td>7.16</td>
<td>Genre families of writing assignments in various disciplines</td>
<td>217</td>
</tr>
<tr>
<td>7.17</td>
<td>Requirements for the use of source materials in disciplinary courses</td>
<td>218</td>
</tr>
<tr>
<td>7.18</td>
<td>Purposes of written assignments in disciplinary courses (based on twenty-seven assignment sheets)</td>
<td>221</td>
</tr>
<tr>
<td>7.19</td>
<td>Purposes of written assignments in disciplinary courses (based on interviews with three content course instructors)</td>
<td>223</td>
</tr>
<tr>
<td>7.20</td>
<td>Rhetorical functions in each assignment in English A</td>
<td>224</td>
</tr>
</tbody>
</table>
9.10 Rhetorical functions in writing assignments in disciplinary courses (based on twenty-seven assignment sheets) 225
9.11 Length of the twenty-seven writing assignments in sixteen disciplines 226
9.12 Summary of assumptions, backing, and degree of support for the extrapolation inference 229
10.1 Research questions motivated by assumptions in the interpretation/use argument 242
10.2 Demographic information of student participants 245
10.3 Validity argument for the web-researching-to-write test 255
11.1 Interpretation/use argument for the academic grammar test 267
11.2 Contingency table showing correspondences between Pass/Not Pass results on EPT writing and academic grammar test results (fall 2012) 280
11.3 Contingency table showing correspondences between Pass/Not Pass results on EPT writing and academic grammar test results (spring 2013) 281
11.4 Summary of warrants, assumptions, backing, and degree of support for the utilization and consequence implication inferences 284
12.1 Comparison of beginning of semester and final exam test scores in four ESL listening classes 308
12.2 The “success rate” in the final exams in two ESL academic writing classes 309
12.3 Descriptive statistics for the raw scores and factor scores of the variables \( N = 239 \) 311
12.4 Model fit indices of the SEM models \( N = 239 \) 312
12.5 Summary of assumptions, backing, and degree of support for the consequence implication inference of the validity argument for the EPT 316
13.1 Summary of inferences and claims in validity arguments with the types of qualitatively oriented and quantitatively oriented assumptions supporting the respective warrants 334
Contributors

Carol A. Chapelle
Iowa State University, United States of America

YunDeok Choi
Sungkyunkwan University, South Korea

Yoo-Ree Chung
Seoul National University, South Korea

Ahmet Dursun
University of Chicago, United States of America

Heesung Jun
Seoul National University, South Korea

Rie Koizumi
Juntendo University, Japan

Hye-won Lee
Cambridge Assessment English, United Kingdom

Jooyoung Lee
Iowa State University, United States of America

Zhi Li
The University of Saskatchewan, Canada

Moonyoung Park
Chinese University of Hong Kong, Hong Kong

Erik Voss
Teachers College, Columbia University, United States of America

Hyejin Yang
Chung-Ang University, South Korea
Series Editor’s Preface

This book, comprising contributions from an international team of researchers, explains and demonstrates the framework of argument-based validity as it applies to language testing.

Language testing is a field of continuing educational and social importance. Language tests enable curriculum designers to meet the needs of learners; they also have a significant gate-keeping role in controlling access to education, jobs, and immigration. It is crucial, therefore, to have accurate information about the extent to which any given test is fit for purpose. The framework of argument-based validity has been applied to language testing since the beginning of the twenty-first century, but it can be difficult for test writers to use, as it requires an explicit statement of the inferences they have made. This book enables test writers to rise to that challenge.

The editors of the book explain the concept and application of argument-based validity with exceptional clarity. The book also provides detailed demonstrations of how it can be used to assess existing tests in a series of chapters describing research projects from several countries. These include tests of spoken English for professional purposes, a variety of writing tasks used to assess readiness for university entry, and an academic collocation test. The methods used in the individual research projects are varied and include both quantitative and qualitative approaches, but each chapter includes an explicit statement of the argument that informs the assessment of validity of the test. Together the chapters add up to a powerful case for the application of argument-based validity to language testing.

This book will be essential reading for anyone with an interest in language testing, whether that is from the perspective of a researcher, a test developer, or a user of test results. It is a valuable addition to the series.

Susan Hunston