

Contents

Preface	viii
1 Introduction	
The aims of the book	1
The climate for language testing	2
Research and development: needs and problems	8
Research and development: an agenda	12
Overview of the book	13
Notes	15
2 Measurement	
Introduction	18
Definition of terms: measurement, test, evaluation	18
Essential measurement qualities	24
Properties of measurement scales	26
Characteristics that limit measurement	30
Steps in measurement	40
Summary	49
Notes	50
Further reading	52
Discussion questions	52
3 Uses of Language Tests	
Introduction	53
Uses of language tests in educational programs	53
Research uses of language tests	67
Features for classifying different types of language test	70
Summary	78
Further reading	79
Discussion questions	79
4 Communicative Language Ability	
Introduction	81

Language proficiency and communicative competence	82
A theoretical framework of communicative language ability	84
Summary	107
Notes	108
Further reading	109
Discussion questions	109
5 Test Methods	
Introduction	111
A framework of test method facets	116
Applications of this framework to language testing	152
Summary	156
Notes	157
Further reading	158
Discussion questions	159
6 Reliability	
Introduction	160
Factors that affect language test scores	163
Classical true score measurement theory	166
Generalizability theory	187
Standard error of measurement: interpreting individual test scores within classical true score and generalizability theory	197
Item response theory	202
Reliability of criterion-referenced test scores	209
Factors that affect reliability estimates	220
Systematic measurement error	222
Summary	226
Notes	227
Further reading	232
Discussion questions	233
7 Validation	
Introduction	236
Reliability and validity revisited	238
Validity as a unitary concept	241
The evidential basis of validity	243
Test bias	271
The consequential or ethical basis of validity	279
Post mortem: face validity	285

Summary	289
Notes	291
Further reading	294
Discussion questions	294
8 Some Persistent Problems and Future Directions	
Introduction	296
Authentic language tests	300
Some future directions	333
A general model for explaining performance on language tests	348
<i>Apologia et prolegomenon</i>	351
Summary	356
Notes	358
Further reading	358
Discussion questions	358
Bibliography	361
Author index	395
Subject index	397

Author index

- Adams, M. L. 45
Alderman, D. L. 277-8
Alderson, J. C. 124, 273, 288, 313, 314, 316, 324, 333
Allen, J. P. B. 86, 115, 352, 354
Angoff, W. H. 344
Austin, J. L. 90
- Bachman, L. F. 5-6, 12, 55, 86, 96, 112, 115, 125, 137, 144, 148, 153, 155, 225, 265, 316, 328, 334, 339, 340, 347, 354, 358 (n 1)
Berk, R. A. 231 (n 28)
Besnier, N. 109 (n 4), 158 (n 6)
Blum-Kulka, S. 310
Bolus, R. E. 195
Bowen, J. D. 118
Brennan, R. L. 116, 213, 224
Brière, E. 273
Brown, G. 131, 133, 138, 158 (n 2)
Brown, J. D. 213
Brown, W. 174
- Campbell, D. T. 225, 239, 263
Canale, M. 85, 99, 296, 316, 317, 321, 351
Candlin, C. N. 84
Carroll, B. J. 329
Carroll, J. B. viii-x, 6, 19, 20, 53 (n 3), 72, 106, 108 (n 2), 116, 117, 187, 255, 301, 302, 308, 345, 353
Cartier, F. 310, 338
Cattell, R. B. 252, 286, 288
Chacevych, A. 268, 273
Chapelle, C. 275, 276, 277
Chavanachart, P. 273
Chen, Z. 273
Chomsky, N. 108 (n 1), 296
Clark, E. V. 100
Clark, H. H. 100
Clark, J. L. D. 5, 12, 69, 116-17, 127, 148, 303, 304, 305, 306, 307, 314, 316, 334, 339-40
Clifford, R. T. 45
Cohen, A. D. 116, 269, 335
Condon, E. C. 273
Cooper, R. L. 310, 355
Corder, S. P. 101
Coulthard, M. 138
- Cronbach, L. J. 115, 158 (n 3), 177, 244, 249, 255, 257, 286, 288, 291 (n 1), 292 (n 6)
Cummins, J. P. 131, 143
Cziko, G. A. 210, 292 (n 7)
- Davidson, F. G. 153, 265
Davies, A. 287, 288, 289
Douglas, D. 10, 113, 324, 335, 337
Duran, R. P. 272
- Ebel, R. L. 181
Erickson, M. 273
- Færch, C. 99, 100, 107
Farhady, H. 278, 279
Feldt, L. S. 116
Firth, J. R. 111-12
Fischer, R. A. 268
Fiske, D. W. 225, 239, 263
Flanagan, J. C. 344
Fouly, K. A. 354
Fraser, B. 91-2
- Gardner, R. C. 69, 354
Glaser, R. 341-2, 343
Greenbaum, S. 158 (n 6)
Griffin, P. E. 154, 307
Grotjahn, R. 155, 270, 335
Guttman, L. 115, 175
- Hagen, E. P. 40
Hale, G. 273
Haley, J. 93
Halliday, M.A.K. 82, 92, 95, 96
Hambleton, R. K. 206, 245
Hamp-Lyons, L. x
Hansen, J. 275
Hansen, L. 276
Hawkey, R. 117, 358 (n 2)
Henning, G. 158 (n 3), 273
Holland, P. W. 277-8
Horst, P. 229
Hoyt, C. 229 (n 12)
Hughes, A. 331
Hulin, C. L. 281, 282, 283, 284
Hymes, D. H. 82, 83, 85, 108 (n 1), 112, 138

396 *Author index*

- Ilyin, D. 278
 Ingram, D. E. 287
- Jackendoff, R. 331
 Jacobs, H. J. 330
 Johnson, K. 102
 Jones, R. L. 304, 305, 307, 311, 314, 315
 Joos, M. 96
- Kachru, B. J. 158 (n 6)
 Kasper, G. 99, 100, 107
 Kenny, D. 260
 Klaus, D. J. 341-2
 Klein-Braley, C. 313
 Kramsch, C. 83, 302
 Krashen, S. D. 133
 Kuder, G. F. 176
 Kunnan, A. 153
- Labov, W. 10, 112
 Lado, R. viii, ix, 82, 302
 LeBlanc, R. 148
 Linnville, S. E. 155, 265
 Livingston, S. A. 221
 Lord, F. M. 230 (n 19), 345
 Lowe, P. Jr. 5
 Lynch, B. 153
- McGroarty, M. 278
 McIntosh, A. 95, 96
 Mack, M. 354
 Mackey, W. F. 116
 Madaus, G. F. 292 (n 2)
 Madsen, H. S. 123, 305
 Meehl, P. E. 255
 Messick, S. A. 236, 238, 247, 249, 255, 269,
 281, 282, 292 (n 6), 309, 324
 Miller, G. 100
 Molloy, J. 273
 Morrow, K. 117, 321, 351
 Mosier, C. I. 285, 323
 Mountford, A. 319
 Munby, J. 84, 117
- Nevo, D. 348
 Nitko, A. J. 283, 340, 342, 345
- Oller, J. W. Jr. 4, 6-7, 10, 106, 124, 239, 288,
 302, 316, 353
 Olshtain, E. 310
 Oxford, R. L. 69
- Painchaud, G. 148
 Palmer, A. S. viii, 86, 96, 105, 113, 115, 125,
 128-9, 137, 144, 148, 225, 268, 347, 354,
 358 (n 1)
 Pawley, A. 97
 Pedhazur, E. J. 293 (n 7)
 Perkins, K. 155, 265
 Peterson, C. R. 310
 Peterson, N. S. 344
 Pike, L. W. 148
 Pilliner, A. E. G. 37, 50 (n 1)
- Plaister, T. H. 273
 Politzer, R. L. 278
 Popham, W. J. 129, 245, 342
 Porter, D. 318-19
 Powers, D. E. 277
- Raatz, U. 51 (n 4)
 Raffaldini, T. 330
 Rea, P. M. 307
 Reves, T. 314, 319, 320
 Richardson, M. W. 176
 Roberts, C. 276, 277
 Roid, G. H. 158 (n 7)
- Sang, F. 354
 Savard, J-G. 116
 Savignon, S. J. 55, 83, 328, 339
 Searle, J. R. 90, 108 (n 3), 140, 313
 Selinker, L. 10, 324, 335, 337
 Shohamy, E. 310, 314, 319, 320, 335, 352
 Skehan, P. x, 311, 312
 Spearman, C. 174
 Spolsky, B. 280, 284, 301, 308, 310, 311,
 313, 314, 318, 320, 331, 332, 353
 Spurling, S. 278
 Stanley, J. C. 116, 163
 Stansfield, C. W. 275
 Stevenson, D. K. 241, 287, 310, 311, 325,
 332, 335, 352
 Strevens, P. 95, 96
 Subkoviak, M. J. 217
 Swain, M. 85, 99, 317, 320
 Swales, J. 95
 Swaminathan, H. 206
 Swinton, S. S. 148, 277
 Syder, F. H. 97
- Takala, S. 246
 Tarone, E. E. 98, 99
 Tatsuoka, K. K. 336, 337
 Tatsuoka, M. M. 336, 337
 Theunissen, T. J. J. 195
 Thorndike, R. L. 40, 116, 163, 344
 Thurstone, L. L. 344-5
- Upshur, J. A. 62, 251, 308
 Urquhart, A. H. 273
- van Dijk, T. A. 82, 89
 van Weeren, J. 195
 Vanniarajan, S. 153
 Venneman, T. 134
 Vincent, R. J. 45
- Wang, L-S. 354
 Weir, C. J. 117
 Wesche, M. 306, 312, 315, 330
 Widdowson, H. G. 10, 87, 141, 316, 319
 Witkin, H. A. 275
- Yule, G. 131, 133, 138
- Zeidner, M. 273

Subject index

- ability 19, 108 (n 1)
ability estimates
 Bayesian approach 345
ability scale
 definition 345
ability score, *see* item response theory
Absolute Language Proficiency Ratings 343
 see also Foreign Service Institute
absolute scales of measurement 344–6
absolute zero point
 measurement scales 26
 ratio scale 29
abstract information 135–6
 see also information
accountability 55
achievement test 60–1, 70, 71
 limited use in summative evaluation 62
 use for developing new programs 62
ACTFL, *see* American Council on the
 Teaching of Foreign Languages
adaptive input and response 150–2
 see also input and response
adaptive testing 150–1
 computer applications 151–2
agreement index 212
 squared-error loss agreement indices
 217–19
 threshold loss agreement indices 217
ambiguity tolerance 277
American Council on the Teaching of Foreign
 Languages oral interview 41
 description of advanced level 41
American Council on the Teaching of Foreign
 Languages Proficiency Guidelines 5, 325,
 344
American Psychological Association
 Standards for Education and
 Psychological Testing 51 (n 4), 236–7,
 243
amplified objectives 154
analysis of variance 193–4
ANOVA, *see* analysis of variance
APA, *see* American Psychological Association
applied linguistics
 challenges for language testing 297
 relation to language testing 352
appraisal
 as synonym for measurement 50 (n 1)
appropriateness of utterances 90
aptitude tests, *see* language aptitude tests
artificiality 322
assessment
 as synonym for measurement 50 (n 1)
attainment test 70
 see also achievement test
Australian Second Language Proficiency
 Ratings 329
authentic test
 lists of characteristics 313
authenticity 9–10, 112, 141, 300–28
 definition 301–3, 316
 effects on test taker's perception of test 151
 effects of test method 112, 318–19
 at heart of language testing 330
 importance of 301
 interactional/ability view 315–23
 as interactive language use 315–23
 real-life approach to defining 301–2
 relation to definition of ability 10
 relation to validity 323–30
 synonymous with functionality 141
 threats to 320
 types of 317–18
Bachman-Palmer oral interview, *see* Oral
 Interview Test of Communicative
 Proficiency in English
Bachman-Palmer scales 326–9
background knowledge
 as source of test bias 273–4
band interpretation 201
band score 201
bias, *see* test bias
BICAL 213 (n 22)
bilingual proficiency 69
Bilingual Syntax Measure 129, 145
Cambridge-TOEFL Comparability Study 152
categorical scales, *see* nominal scales
Certificate of Proficiency in English 120–1
 compared to TOEFL 153
 familiarity with test procedures 114
 use of British English 143
characteristics
 as part of measurement 19–20
CLA, *see* communicative language ability
classical test theory, *see* classical true score
 theory

- classical true score theory
 approaches to reliability 184–5
 compared to G-theory 196–7
 error variance treated as homogeneous 186–7
 operational definition of ability 210
 problems with model 185–7
 as a special case of G-theory 188
see also true score
- classification errors
 differences in seriousness 219–20
 differential costs of 216
 loss ratio 216
 with mastery/nonmastery classification decisions 215–16
- clichés
 interpretation of 98
- cloze test 48–9, 76, 114
 computer administered 337
 criteria for correctness 124
 effect of field dependence 164, 275–6
 factor analysis of 265
 research on facets 155
 strategies of test takers 121, 269
 use in research 68
- coefficient alpha 177
 coefficients of reliability, *see* reliability
- cognitive characteristics
 as source of test bias 274–5
- cohesion 88–9
 scales of ability 327 (8.1b)
- common metric scale 5–6, 75
 essential characteristics 340–4
 need for developing by criterion-referenced tests 338–40, 348
- communalities 262
- communication
 characterization of 83
 as dynamic interchange between context and discourse 102
 minimal requirements of measures 316
- communication strategies
 interactional definition 98–9
 psycholinguistic definition 98, 100
- communicative competence 16 (n 2)
 research into models of 68
- communicative goals
 elements of 100
- communicative interaction 83
- communicative language ability 4
 in communicative language use 85 (4.1)
 components of 84
 defined 14
 described 84
 development of criterion-referenced measures for 334
 expanded frameworks for 334–5
 as label for language ability 243
 necessity for authentic testing methods 325
 validation of authentic tests of 334–8
 variation from one test task to another 113
see also communicative proficiency;
 communicative language proficiency
- communicative language proficiency 4
see also communicative proficiency;
 communicative language ability
- communicative language tests
 desirable characteristics 320
- communicative language use
 distinguishing feature of 316
- communicative proficiency 4, 36
see also communicative proficiency;
 communicative language ability
- communicative test events 117
- comparability
 of language tasks 36
- competence
 versus knowledge 108 (n 1)
 versus performance 52 (n 8)
- complete language ability 38
- comprehensibility
 of language test input 319
- computer adaptive tests 37, 121
- concrete information 135–6
see also information
- concurrent criterion relatedness, *see* concurrent validity
- concurrent criterion relevance, *see* concurrent validity
- concurrent validity 248–50
- confidence intervals
 for standard error of measurement 200–1
- confirmatory factor analysis, *see* factor analysis
- confirmatory mode
 of correlational approach to construct validation 260
- construct
 basis for operational definitions 47–8
 defined 255
 defining theoretically 41–2
 operational definition 42–4, 48
 relation between theoretical and operational definitions 46
 as synonym for competence 108 (n 1)
 as a way of classifying behavior 255
- construct definition 155
- construct validation 6–7, 254–6
 benefits of research 333
 contrasted with experimental design 267
- correlational approaches 260–2
 as empirically testing hypothesized relationships between test scores and abilities 256
 experimental approach 266–8
 requires logical analysis 256–7
 tasks involved 270–1
 types of empirical evidence collected 258
- construct validity 115
 ethical use and interpretation 281
 as function of authenticity 323–4
 in interactional/ability approach 303
 variation across individuals 324
- content coverage, *see* content validity
- content domain
 definition 48

- content-referenced interpretation 231 (n 27)
- content relevance, *see* content validity
- content representativeness, *see* content validity
- content validity
- content coverage 245, 311
 - content relevance 244
 - content representativeness 306–7
 - of direct proficiency tests 306–7
 - primary limitation 247
- context
- interaction with discourse 82–3
 - of language use 82
 - relationship to assessment component of strategic competence 100–1
- context embeddedness 131
- context of situation 112
- contextual information 131–4
- contextualization
- degree of 131–4
- convergence
- among correlations 263
- conversational competence 88
- conversational language 88–9, 146
- conversational routines 88
- conversation rules 88
- copytest 337
- correlation
- defined 259
- correlational analyses
- applications 265
- correlation coefficient
- defined 259
 - used in supporting or rejecting interpretations of test scores 259–60
- correlational evidence
- for construct validity 259–65
- correlational inference 260
- co-text 133
- counterbalanced test design 183 (6.2)
- counterfactual information 136–7
- see also* information
- CPE, *see* Certificate of Proficiency in English
- CR, *see* criterion-referenced
- criterion ability
- measuring limited aspects of 251
- criterion-referenced agreement indices
- sensitivity to difference in cut-off score 221
- criterion-referenced approach
- basis for common metric scale of language 340
- criterion-referenced measurement 7–8
- criterion-referenced standard error of measurement 213–14
- criterion-referenced test 59, 72, 74–6
- categories 341 (8.1)
 - defined 342
 - development of, with specified domains 343
 - as distinguished from norm-referenced test 75–6
 - effects of score variance on reliability coefficients 221
 - estimating reliability 219–20
 - interpretation using standard error of measurement 213–14, 219–20
 - problems with domain specification 338
- criterion-referenced testing 8, 154
- criterion-referenced test scores
- contextual applicability 211
 - interpretation 210
 - reliability 209–17
- criterion relatedness, *see* criterion validity
- criterion validity 248–54
- see also* concurrent criterion validity; predictive validity
- Cronbach's alpha, *see* coefficient alpha
- C-test 270
- CTS (classical true score), *see* classical true score theory
- cultural content
- as source of test bias 272–3
- curricular validity 292 (n 2)
- cut-off score 214–15
- effect on criterion-referenced agreement indices 221
- DBP Study, *see* Development of Bilingual Proficiency: Second Year Report
- decision errors 56–7
- minimization 202
 - see also* classification errors
- decision making
- two components of 284–5
- decision study 188
- dependability coefficient 212
- dependability index 212–13
- developmental sequence, *see* second language acquisition
- Development of Bilingual Proficiency: Second Year Report 353, 354
- diagnostic test 60, 70
- dialect 142–3
- sensitivity to differences in 95
 - standard 142–3
- dichotomous scale 27
- dictation 48, 76
- difficulty parameter 204–5
- dilemma of language testing 2, 9, 287–8
- direct test 287, 309
- discourse
- interaction with context 82–3
 - discourse, field of, *see* field of discourse
 - discourse, mode of, *see* mode of discourse
 - discourse, oral, *see* oral discourse
 - discourse, style of, *see* style of discourse
 - discourse analysis 88
 - discourse community 96
 - discourse competence 85
 - discourse domain 10, 95–6, 324, 337
 - discourse grammar 85
 - discourse level of operation 84
 - discrete-point language test 34, 128
 - discrimination
 - among correlations 263 - discrimination parameter 204–5

- distinctiveness
 in interval scales 28, 30
 in measurement scales 26, 30
 in nominal scales 27, 30
 in ordinal scales 28, 30
- domain
 ordered 342
 types of 341
 as universes of admissible observations 246
 unordered 342
- domain of content 75
- domain of language ability
 difficulties in defining 245
 problems in specification 246, 311
- domain-referenced 72
- domain score dependability, *see* dependability
- domain score estimates, *see* dependability
- domain specification, *see* domain of language ability
- D-study, *see* decision study
- educational programs
 decisions about teachers 61
- Educational Testing Service
 EFL proficiency test batteries 152
- edumetric tests 75
- elicitation procedure 9
 effect on language learners' interlanguage system 113
 effects on test performance 21
 similarities in IA and RL approaches 331
see also test method
- elicitation task
 with separate rating scales 330
 English for specific purposes testing 273-4
 English Language Testing Service 121, 320
 test of 47, 74
- entrance test 70
see also selection test; readiness test
- equal intervals
 in interval scales 28
 in measurement scales 26
- equivalence 182-4
see also reliability
- error analysis
 content based 336
- error score 167, 228 (n 2)
 as factor affecting observed score 167 (6.2)
 for individuals 199
- error score variance
 defined in terms of total score variance and reliability 171
 in classical measurement theory 199
 treated as homogeneous with CTS model 186
- errors in classification decisions, *see* mastery/nonmastery classifications
- errors of measurement, *see* measurement error
- ESP testing, *see* English for Specific Purposes testing
- ethical values
 and validity 237
- ethics of observation 314
- ethics of test use 280
- evaluation
 components of 54
 defined 22
 information provided by 54-5
 non-test measures for 23-4
 relationship with measurement and tests 23 (2.1)
see also formative evaluation; summative evaluation
- examination
 as distinct from test 50-1 (n 1)
- expected response 125-48
 format of 129, 130
 form of 130
 genre 138-9
 language of 130
 restrictions on channel 144-5
 restrictions on format 145
 restrictions on grammatical forms 145
 restrictions on organizational characteristics 145
 restrictions on propositional and illocutionary characteristics 146-7
 restrictions on time or length 147-8
 topic 137-8
 type of 129, 135
see also input; input and expected response; response
- experimental design 266-8
- exploratory factor analysis, *see* factor analysis
- exploratory mode
 of correlational approach to construct validation 260
- face/content validity 323
- facet theory 115, 157 (n 1)
- face validity 285-9, 302, 303, 323
 criticisms of 285-9
 defined in terms of authentic test language 315
 of direct proficiency tests 306
 rejected as basis for justifying test interpretation 310
 three meanings of 285-6
- factor analysis 262-3
 confirmatory 262, 263-4, 265
 exploratory 262, 265
- factor loadings 262
- FCE, *see* First Certificate in English
- feedback 55, 149, 151
- field independence
 as factor influencing language acquisition 275-6
 as predictor of performance on tests of English 276
- field of discourse 94
 defined 95

- figurative language
 interpretation of 97–8
 First Certificate in English
 compared to TOEFL 153
 Foreign Service Institute 344
 Foreign Service Institute scale 45
 see also ILR scale
 Foreign Service Institute-type scale definition
 344
 formative evaluation 60
 see also evaluation
 F ratio 168
- GEFT, *see* Group Embedded Figures Test
 generalizability coefficients 192, 194–5, 196
 generalizability study 188
 example with placement test 190
 generalizability theory 7, 162, 187–97
 advantages over classical test theory 196–7
 population of persons 190–1
 practical application to language testing
 195–6
 universe score 191–2
 general factor of language proficiency 106
 general language proficiency 6
 GENOVA 230 (n 18)
 genre 138–9
 effect on discourse interpretation 138
 genuine validity 306
 g-factor 6
 grammatical competence 85, 86, 87,
 326–7 (8.1a)
 examples 92
 Group Embedded Figures Test 275
 G-study, *see* generalizability study
 G-theory, *see* generalizability theory
 guessing parameter 204–5
 GUME study 354
 Guttman split-half estimate 175, 177–8, 210
- heuristic function 93–4
 hyperboles
 interpretation of 98
 hypotheses 256–8
- IA, *see* interactional/ability
 ICC, *see* item characteristic curve
 ideational function 92–3
 IL research, *see* interlanguage research
 Illinois Study 321
 illocutionary act 90
 illocutionary competence 42, 90–4
 examples 91–2
 illocutionary force 90, 140–2
 ILR, *see* Interagency Language Roundtable
 Ilyin Oral Interview 129
 imaginative function 94
 imprecision
 of language tests 35
 incompleteness
 of language tests 33–5
 indirectness
 of language tests 32–3
- information
 concrete/abstract 135–6
 counterfactual 136–7
 negative 136
 information functions 207–9
 information gap 134, 320
 input 117, 119 (5.1), 125–44
 channel and mode of presentation 127
 defined 125
 degree of contextualization 131–4
 degree of speededness 129–30
 dialect 142–3
 distribution of new information 134–5
 format of 127
 form of presentation 127
 genre 138–9
 identification of the problem 128
 illocutionary force 140–2
 language of presentation 127
 length of language samples 130
 organizational characteristics 140
 register 143–4
 sociolinguistic characteristics 142
 type of information 135–7
 vehicle of presentation 127
 vocabulary 131
 see also expected response; input and
 expected response; response
 input and expected response
 adaptive, *see* adaptive input and response
 characteristics of relationships between
 151 (5.1)
 issue for authenticity 319
 nonreciprocal, *see* nonreciprocal input and
 response
 reciprocal, *see* reciprocal input and
 response
 see also expected response; input;
 response; language test
 input component
 of language testing 335, 352
 instrumental function 93
 instructional validity 292 (n 2)
 intelligence
 relationship to language abilities 106
 interactional/ability approach 15, 42, 302–3
 authenticity 317–22, 324–6
 compared with real life approach 325–30
 difficulties in use 331
 see also authenticity
 interactional competence 302
 interactional function 93
 Interagency Language Roundtable 4–5
 Interagency Language Roundtable Language
 Skill Level Descriptions 325
 Interagency Language Roundtable oral
 interview 21, 41, 44, 120, 129, 302
 interlanguage research 324
 internal consistency 172–8, 184
 effects of violating assumptions 178 (6.1)
 see also split-half reliability; Kuder-
 Richardson reliability coefficient;
 coefficient alpha

- inter-rater consistency 180-1
- inter-rater reliability 180-1
- interval scale 28-9
 - comparison with ordinal scale 28 (2.2)
- intra-class correlation coefficient 229 (n 12)
- intra-rater reliability 178-80
- IRT, *see* item-response theory
- item characteristic curve 203-6
 - and item information function 207
- item characteristics
 - types of information 204
- item forms 154
- item information functions 207-8, 208 (6.7)
- item-response theory 7, 37, 162, 202-9, 265
 - ability score 206
 - advantages over classical true score theory 209
 - item information function 207-8
 - one parameter model, *see* Rasch model
 - test information function 208-9
 - three parameter model 204-5
 - two parameter model 205
 - unidimensionality assumption 203
- item response theory models
 - application considerations 204
- job analyses
 - use in domain specification 311-12
- knowledge
 - synonym for competence 108 (n 1)
- knowledge acquisition continuum 342
- KR-20, *see* Kuder-Richardson formula 20
- KR-21, *see* Kuder-Richardson formula 21
- Kuder-Richardson formula 20 176, 210
- Kuder-Richardson formula 21 176
- Kuder-Richardson reliability coefficients 176
- language
 - as instrument and object of measurement 2, 287-8
- language ability 16 (n 2)
 - definition 3-4
 - stages of development in the history of research in 353-4
 - labels for 243
 - problem of specification of 8, 31
 - two ways of viewing 251-4
 - versus intelligence 106-7
 - versus language performance 308-9
 - views of 4-5
- language acquisition 15-16 (n 1)
 - effects of instruction settings and techniques 69
 - relation to language testing 2-3
 - use of language tests for 68-9
- language aptitude tests 72
 - language attrition 69
 - language competence 84-7
 - components of 87 (4.2)
- language education programs
 - evaluation of processes 55
 - formative evaluation of 62
 - measurable outcomes 55
 - summative evaluation of 62
- language functions 92-4
 - see also* heuristic function; ideational function
- language instructional programs
 - remedial component 65-7
- Language Monitoring Project 330
- language norms
 - of actual usage 39
 - kind of language chosen for 39
- language performance
 - obtained under uniform conditions 43-4
 - sampling of 33-5
 - versus language ability 308-9
- language proficiency 16 (n 2)
 - consists of several distinct but related constructs 68
 - contrasting views 5, 251-4, 254 (7.2)
 - definition of criterion-referenced scales 326-8 (8.1), 346-8
 - development of common metric of 339-40
 - differences between real life and interactional/ability definitions 329
 - as real life performance 41, 303
 - research agenda for 340, 354-5
 - skills/components model 82
- language proficiency test
 - problems in development 47
- language proficiency theory
 - effect on test development 47
- language programs 63 (3.1), 64 (3.2), 65 (3.3), 66 (3.4)
 - evaluation of 339
- language teacher
 - decisions about 61
- language teaching
 - communicative view 3
 - relation to language testing 2-3
- language test
 - assumptions for use in educational programs 55
 - based on syllabus used 71
 - based on theory of language proficiency 71
 - benefits of computer administration 336
 - cataloging of 116
 - classifying types of 70-8
 - comparative description of 152
 - content of 71-2
 - context in which takes place 43
 - criteria for correctness 124-5
 - design 9, 153-4
 - functionality of 141-2
 - difficulty in identifying language ability measured by individual items 174
 - dissatisfaction with 351
 - domain-referenced interpretation 72, 343
 - educational and social consequences 237
 - ethical considerations 57, 284-5
 - factors affecting performance 163-6
 - features of 70-8
 - frame of reference 72
 - instructions 123-5
 - as instruments of social policy 280
 - intended use 70-1
 - key measurement problem 11

- length of samples 130
- necessity to examine process as well as product 335
- as operational definitions of theoretical constructs 67-8
- practical usefulness 282
- prediction as one use of 253
- problems in comparison 152
- provision of diagnostic information 60
- purposes served 54
- quality of information provided 56
- research uses of 67-9
- scoring procedure 76
- specification of procedures and tasks 123-4
- time allocation 122-3
- use for diagnosis 60
- use for placement 58-60
- use for progress and grading 60-1
- use for selection 58
- use in decisions about programs 62-7
- use in determining abilities of language learners 253
- use in developing tests of constructs 68
- use in educational programs 54-67
- use in research about nature of language processing 68
- use in research of language acquisition 68-9
 - as valid measure of abilities affected by a treatment 267
- language test content
 - identifying abilities 40
- language testers
 - collaboration with consumers of language tests 355
- language testing
 - areas of current research 2
 - challenges to 296-300
 - conceptual foundation 1-2
 - conferences on xi
 - context 2
 - convergence of two disciplines 296-9, 352-5
 - discrete structure-point approach 300
 - educational context 62-7
 - general consideration for use 56
 - historical development of 296-7
 - input component 352
 - integrative approach 300
 - integrative-sociolinguistic trend 299
 - interdisciplinary nature of 297
 - output component 352
 - psychometric-communicative trend 299
 - psychometric principles 352
 - psychometric-structuralist trend 299
 - publications for xi
 - relation to language acquisition research and language teaching 2-3
 - research agenda 12-13
 - sociolinguistic principles 352
- language testing research
 - challenges to 352-3
 - future directions 333-48
 - goals 13, 155
- language test performance
 - affected by individual attributes 113-14, 277-9
 - effect of cognitive characteristics 274-7
 - effect of cultural context 272-3
 - effect of prior knowledge of content 132, 273-4
 - factors that affect 163-6
 - general model 348-50
 - need for theoretical framework of affecting factors 316
 - variation across different testing situations 113
- language test research
 - highest priority 334
- language test scores
 - factors that affect 163-6, 348-9
 - sources of variation 350 (8.2)
- language use
 - defined 83
 - model of 103 (4.3)
- language use tasks
 - identification of authentic 332
- latent trait theory, *see* item response theory
- length of test
 - reliability differences 220
- levels of measurement, *see* scales of measurement
- lexical competence 87, 97
- limitations in observation and quantification 32-40
- limitations in specification 30-2
- limitations on measures 30-40
- linguistic encoding 84
- logical task analysis 155, 270
- LOGIST 231 (n 22)
- macro-evaluation 58
- manipulative functions 93
- mastery level
 - as domain score indicative of minimal competence 214-15
- mastery/nonmastery classifications
 - dependability of 214-20
 - errors in 214-15
 - estimating dependability 216-20
- mastery test 61, 70
- mean
 - defined 166
 - use in computing reliability coefficients 176
- measure
 - ambiguity of the term 51 (n 3)
- measurement
 - definition 18-19
 - impressionistic approaches 50 (n 1)
 - operationist approach 309
 - relationships to tests and evaluation 23 (2.1)
 - use in social sciences 51 (n 2)
- measurement error 167, 170-2
- considered random by CTS model 187
- estimates at each ability level 208
- random 164
- and reliability 24

- measurement error (cont.)
 - sources of 161–4
 - systematic 164, 187, 222–6
 - see also* reliability; standard error of measurement
- measurement instruments
 - distinguished from test 20–1
- measurement models
 - applications to analysis of language test scores 298
- measurement scales, *see* scales of measurement
- measurement steps 40–9
 - relevance for development of language tests 45
 - relevance to interpretation of test results 48–9
- measurement theory 11
 - applications to language testing 6–8, 296–9, 352–3
- measurement units
 - definition of 44–5
- metalinguistic response 129
- metaphors 98
- micro-evaluation 58
- microscale 231 (n 22)
- minimum competency testing 214, 232 (n 30), 338–9
- mode of discourse 95
- Modern Language Aptitude Test 72
- Modern Language Centre of the Ontario Institute for Studies in Education 69
- MTMM, *see* multitrait-multimethod
- multiple-choice 115–16
- multiple-choice reading test
 - strategies of test takers 269
- multiple regression analysis
 - application in EFL proficiency test 265
- multitrait-multimethod correlation matrices
 - analysis of 263–5
- multitrait-multimethod design 263–5
 - relationship among traits, methods and measures 264
- native language background
 - as source of bias in language tests 277
- native-like way 97
- native speaker
 - applied linguistics view of 343–4
 - performance of 248–9
- native speaker norms 38–40, 52 (n 9)
- negative information 136
 - see also* information
- nominal scale 27
- noise test experiments 145
- nonreciprocal input and response 149–50
 - see also* input and response
- nonreciprocal language use
 - defined 149–50
- nonreciprocal language performance
 - 150 (5.3)
- nonreciprocal test tasks 150
- normal distribution 72–3, 73 (3.6)
- norm group 72
- norm-referenced approach to measurement
 - inadequate for developing common measures 340
- norm-referenced reliability
 - as special case of criterion-referenced reliability 221
- norm-referenced test 59, 72–4
 - as distinguished from criterion-referenced test 75–6
 - interpretation 72–4
- norm-referenced testing 7–8
- norm-referenced test scores
 - contextual applicability 211
 - interpretation 210
- NR, *see* norm-referenced
- objective test 76
- observed score
 - relationship to true and error score 167 (6.2), 169 (6.3)
 - factors affecting 165 (6.1), 167 (6.2)
 - as indicators of domain scores 212, 219
- observed score variance 192
- odd-even method 173
- one parameter model 205
- Ontario Test of English as a Second Language 121, 320
- operational definitions 42–4
 - related to interpretation of test results 48–9
 - related to test development 46–7
 - relationships with theoretical definitions and test scores 46 (2.3)
- OPI, *see* Oral Proficiency Interview
- oral communication test
 - with pictures 146
- oral discourse 88
- oral interview 5, 76, 115
 - effects on performance 111
 - as example of performance test 77
 - authentic conversation example 321
 - measuring consistency 185–6
 - as reciprocal 151
 - reliability and validity 241
 - sources of error 184
- Oral Interview Test of Communicative Proficiency in English 129, 325–9
- oral proficiency
 - interactive/ability approach 325–30
 - as label for language ability 243
 - real life approach 325–30
- Oral Proficiency Interview 325–9
- ordered in magnitude (measurement scales) 26
- ordering
 - in interval scales 28
 - in ordinal scales 28
- ordinal scale 28
 - comparison with interval scales 28 (2.2)
- organizational competence 42, 87–9
- parallel forms reliability 182–3
- parallel tests
 - correlations between ones not experimentally independent 170 (6.4)

- definition 168
- use in defining reliability 168-9
- passage effect 138
- path diagram 164-5
- Pennsylvania Study 354
- performance
 - versus competence 52 (n 8)
- performance style 85-6
- performance test 77, 225
 - definition 304-5
 - degrees of directness 305
- perlocutionary acts 90
- perlocutionary effects
 - of language competencies 90-1
- phatic language 93
- Pimsleur Language Aptitude Battery 72
- placement test 70
 - based on content objectives of the program 59
 - based on theory of language proficiency 59
 - bases for 59
 - as example of application of generalizability theory 189
 - regarded as broad-band diagnostic test 60
 - relative stability of enrollments as a factor in development and use 59-60
- planted encounter 314
- post-test
 - in experimental designs 267
- power test 120, 123
- pragmatic characteristics of tests 140-5
- pragmatic competence 42, 86, 89-98, 256
 - as example of steps in measurement 45-6
 - rating scale 327 (8.1b)
- pragmatic expectancy grammar 4, 302
- pragmatic mapping 4, 106
- pragmatics
 - two aspects of 89
- pragmatic test
 - similar to authenticity 316-17
- precision of measurement 207-9
- prediction
 - role of indeterminacy 252
- predictive utility, *see* predictive validity
- predictive validity 250-4, 302
 - consideration of abilities measured 250-1
 - of direct proficiency tests 306
- prescriptive norms 39
- production strategy 99
- proficiency 16 (n 2)
 - in criterion-referenced scales 341-2
- proficiency test 71
- proficiency testing
 - direct and indirect tests 304
- progress test 70
- propositional act 90
- propositional content of tests 130-9
- pseudo-chance parameter 204-5
- psychometric analysis 299
- psychometric theory, *see* measurement theory
- psychometric test 74
- psychophysiological mechanisms 84, 107
- quantification 19
- quantifying observations 44-5
- random error
 - different from systematic error 223
 - reduced by standardizing test method facets 224
- randomization 266
- Rasch model 205
- rating scale
 - development of 36, 44-5
 - precision 36-7
- ratio scale 29-30
- readiness test 58, 70
 - use of 60
- reading comprehension test
 - focused on one discipline 224-5
 - requiring inference 10, 105
- real life approach 41-2
 - accommodations 313-15
 - compared with interactional/ability approach 317, 325-30
 - criticisms of 308-12
 - disadvantages of 330-1
 - three basic tenets of 303
 - as useful for guiding practical test development 330
 - validity of direct tests 305-7
- real-life language use 9-10
 - ways of observing 314
- real-life performance 301
- reciprocal input and response 148-9
 - see also* input and response
- reciprocal language
 - defined 148-9
- reciprocal language performance model 149 (5.2)
- reciprocal language use
 - interaction as characteristic 149
- reciprocal tests 150
- register 95
 - differences between written and spoken 95
 - sensitivity to differences in 95-7
- regulatory function 93
- relativeness
 - of levels of performance 38
- reliability
 - affected by cut-off score 221
 - affected by length of test 220
 - affected by level of difficulty 220-1
 - affected by test score variance 220-1
 - classical true score approach 184-5
 - coefficients of 172
 - costs involved in assuring 57
 - defined 24
 - defined using proportion of observed score variance to true score variance 170-1
 - defined using scores on parallel tests 169
 - distinguished from validity 160-2, 239-41
 - equivalence 182-3
 - as a matter of generalizability 188
 - parallel forms 182-3
 - practical considerations in the estimation of 209

- as quality of test scores 24
- relation to quantification of observations 49
- relationship with validity 240 (7.1)
- as reliable variance 239
- as requirement for validity 160, 238–9
- remedial instruction 65–7
- representational validity 306
- requesting
 - strategies for 91
- residual variance 193
- resource grammar 85
- response 116–17
 - actual response 125–6
 - definition 125–6
 - constructed response 129
 - degree of contextualization 131–4
 - dialect 142–3
 - distribution of new information 134–5
 - expected response, *see* expected response
 - illocutionary force 140–2
 - length of language sample 130
 - organizational character 139–40
 - register 143–4
 - selected response 129
 - sociolinguistic characteristics 142
 - vocabulary 131
- response, reciprocal, *see* reciprocal response
- rhetorical organization 88
- RL, *see* real life
- rules and procedures (measurement) 20
- rule-space model 336–7
- sampling
 - of tasks from domain 311
- scale calibration 345–6
 - end points 346–7
 - research 347–8
- scales of measurement 26–30
- scales of measurement, absolute, *see* absolute scales of measurement
- scoring
 - as influence of strategic competence on test performance 105–6
- second language acquisition 2–3, 69, 339
 - developmental sequence 3, 339
- selected response 129
- selection tests 70
- self-ratings 148
 - as indicators of language abilities 148
- self-weighting 122
- SEM, *see* standard error measurement
- semi-direct test 127
- sexism in language 16–17 (n 3)
- similes
 - interpretation 98
- SLA, *see* second language acquisition
- sociocultural orientation 84
- sociolinguistic competence 42, 85, 87 (4.2), 94–8
 - scale 328 (8.1c)
- sociolinguistic principle
 - of language testing research 335
- sociosemantic basis of linguistic knowledge 84
- sources of error
 - and approaches to estimating reliability 171–2
 - examination under classical test theory 197
 - examination under generalizability theory 197
 - see also* measurement error
- SPEAK, *see* Speaking Proficiency in English Assessment Kit
- Speaking Proficiency in English Assessment Kit 335–6
- Spearman-Brown prophecy formula 174, 229 (n 8)
- Spearman-Brown split-half estimate 174–5, 177–8
- speech acts
 - theory of 90
- speech events
 - language used in different 112
- speech production
 - psycholinguistic model 100
- speededness 128–9
- speeded tests 121, 123
- split-half reliability coefficient 174–5
 - see also* Spearman-Brown split-half; Guttman split-half
- split-half reliability estimates 172–5
 - difficulties in splitting tests 173–4
- squared-error loss agreement indices 217–19
- stability 181–2, 184
- standard deviation
 - defined 106, 166
- standard error of measurement 198–202
 - and band interpretations in criterion-referenced measurement 219–20
 - confidence intervals 200 (6.5)
 - criterion-referenced 213–14
 - estimation from theorems of classical measurement model 199–200
 - group specific estimate 214
 - for individual test takers 213–4
- standardization of test method 224
- standardized test 74
- standard setting 232 (n 30)
- Standards for Educational and Psychological Testing definition of validity 236
 - on face validity 286–7
- statistical analysis 299
- stimulus 116–17
- strategic competence 84, 98–107
 - assessment component 100–1
 - definition 99–100
 - execution component 103–4
 - function of 102
 - influence on language test performance 104–6
 - measurement of 106–7
 - planning component 101–2
- style of discourse 96
- subjective measure 51 (n 4)
- subjective test 76
- subjectivity
 - of language tests 37–8

- summative evaluation 61, 62
 language proficiency measures needed 62
see also evaluation
- syllabus
 basis for test content 71 (3.5)
 learning objectives as theoretical
 definitions of abilities 46
- systematic error 164, 187, 222–6
 defined within context of generalizability
 theory 222
 different from random error 223
 general effect 222
 introduced by standardizing test method
 facets 224
 specific effect 222
see also measurement error
- tables of specifications 154
- tailored test 150–1
- task performance
 strategies and styles 269–71
- teaching methods
 relation to testing methods 47
- test
 definition 20–2
 as distinct from examination 50–1 (n 1)
 distinguished from measurement
 instrument 20–1
 elicitation of specific behaviors 22
 as means of identifying merit 280, 284
 relationships to measurement and
 evaluation 23 (2.1)
- test appeal 287–8
- test authenticity, *see* authenticity
- test bias 138, 166, 271–9
 and differences in group performance 271–
 2
 forms of 272
- test consistency, *see* reliability
- test content 71–2
 defining 34–5
 as important part of test development and
 use 244
 syllabus-based versus theory-based 71 (3.5)
- test design 153–4
 elements to be specified 245
 minimize effects of attributes not part of
 language ability 166
- test development
 application of generalizability theory 188
 ethical considerations 281–2
 goal of 13
 measurement steps in 45–8
- Test in English for Educational Purposes 320
- test information function 208–9
- testing
 alternatives as a means of achieving the
 same purpose 284
- testing environment 118
 effect on authenticity 318
- test interpretation
 consequential basis for 243
 effects of test method facets 224
 evidential basis for 243
- test items
 local independence 11, 203, 229 (n 6)
- test method
 applied to existing needs 331
 effects of standardization 223–4
 effects on test performance 12, 113, 223–4
 identification as source of measurement
 error 160–2
 relation to teaching methods 47
 systematic facets 164
 types 76–7, 115
- test method facets 111, 115, 116, 119 (5.1),
 164
 effects on language test scores 258
 five major categories 117–18, 119 (5.1)
 use in test design 154
- test modalities 116
- Test of Communicative Competence in
 English 120
- Test of English as a Foreign Language 47, 58,
 73, 74, 121
 compared to CPE 153
 compared to FCE 153
 familiarity with test procedures 114
 re-examination of 332
 use of American English 143
 use of three parameter model 205
- test performance 111
 affected by factors other than language
 ability 164
 affected by weighting of parts 122
 effects of multiple background
 characteristics on 278
 effects of random factors 164
 related to individual's level of ability 203
 sources of error 184
- test preparation 114
- test results
 contrasting views of uses and
 interpretation 254 (7.2)
 four areas to be examined for ethical use
 281
 imperfections of 30
 interpretation 35, 48
- test-retest estimates
 affected by changes in ability 186–7
 error variance 186–7
see also stability
- test-retest reliability 181–2
- test rubric 118, 120–5
- test scores
 accuracy for individuals 197–202
 ambiguity of inferences from 261 (7.3)
 defining standards for interpretation 40
 ethical use of 280–1
 factors affecting 165 (6.1)
 interpreted in generalizability theory 187–8
 limitations on interpretation 30–40
 limitations in specification 30–2
 measurement error for individuals
 197–202
 relationships to theoretical and operational
 definitions 46 (2.3)
 unidimensional 11

- test takers 10
 consideration of rights of 280-1
 explanations for unsuccessful performance 246
 as greatest source of subjectivity 38
- test taking
 eliciting strategies employed 335-6
 knowledge gained from computer administration 336-7
 knowledge gained from using error analysis 336
 processes involved 269-71
 research in 68
 strategies 114, 268-71
- testing technique
 as facet in universe of generalizations 189
- test use
 consequences to education system or society 283-4
 for employment decisions 282-3
 effect on instruction 283
 informed by value systems 281-2
 legal decisions affecting 282-3
 multiple branching approach 331-2
 types of evidence to support 243
- test wiseness 114
- text
 conceptual difficulty in processing 158 (n 2)
- textual competence 88-9
- TIF, *see* test information function
- theoretical definitions
 relationships to operational definitions and test scores 46 (2.3)
- three-parameter item characteristic curves 204 (6.6)
- three-parameter model
 of IRT 204-5
- threshold loss agreement indices 217
- TOEFL, *see* Test of English as a Foreign Language
- topastic error 187
- topic of input 137-8
- trait
 as synonym for construct 108 (n 1)
- true score 167, 191, 228 (n 2)
 measurement theory assumptions 166-7, 228 (nn 1-2)
 relationship to observed score 167 (6.2), 169 (6.3)
- true score model, *see* classical true score theory
- t* test 168
- turn taking, *see* conversational language
- two parameter model 205
- underspecification
 of factors affecting test scores 32
- unitary trait hypothesis 6
- universe
 use of term in G-theory 230 (n 16)
 universe of generalization 189
 universe of possible measures 189
 example of facets and conditions defining 190 (6.3)
- universe score 191-2
- universe score variance 192
- University of Cambridge Local Examinations
 Syndicate tests of 152
see also CPE, FCE
- Utah Study 321
- utterance act 90
- validation
 as broad based process 238
 description by Cronbach 244
 inadequate claims for 309-12
- validity
 by assumption 285-6
 authenticity 323-30
 consequential basis of 279-85
 construct validation as evidential basis for 249
 contrasting views of 254 (7.2)
 costs involved in assuring 57-8
 definitions 25, 236-7
 distinguished from reliability 160-2, 239-41
 effect of ethical values 237
 effects of systematic error 223
 ethical basis of 279-85
 evidential basis of 243-71
 facets of 242 (7.1)
 progressive matrix for 242 (7.1)
 as quality of test interpretation and use 25
 relationship to reliability 240 (7.1)
 and reliable variance 239
 types of evidence 237
 unified framework of 242
 as a unitary concept 241-3
- validity, construct, *see* construct validity
- variance
 defined 166
 sources of 192-4, 349, 350 (8.2)
 use in computing reliability coefficients 175, 176, 177, 180
- variance components 188
 estimates using ANOVA 193-5
- vocabulary 131
 scales of ability 327 (8.1b)
- A Vous la Parole* 320
- washback 283
- z* test 168