

Index

A	
<i>a posteriori</i> comparisons	250–52, 256
<i>a priori</i> comparisons	250–51, 256
<i>a priori</i> grouping	291, 292
abscissa <i>see</i> horizontal axis	
absolute decisions	
absolute error variance	186–88
certification/selection decisions	11
criterion-referenced tests	31
dependability	188–89, 192
standard error of measurement	195, 196, 205
test score reporting	294, 300, 321
<i>see also</i> relative decisions	
accuracy	
accuracy/confidence trade-off	235
error of prediction/estimation	107
measurement procedures	28–29
rounding, of decimals	57n, 86, 89
sample statistics	35–36
standard deviation, checking for reasonableness of	69
achievement tests	
criterion-referenced tests	193
domain-referenced tests	31
percentile ranks	306
agreement indices	
coefficient kappa ($\hat{\kappa}$)	201–2
estimated proportion of agreement (\hat{p}_o)	200–202
kappa squared ($\kappa^2_{(X,T_X)}$)	203–4
phi lambda (Φ_λ)	203–4
squared-error loss agreement indices	203–5
threshold loss agreement indices	200–202
Alderson, J.C.	4, 9, 15, 17, 93, 118, 120, 137, 275, 301
Almond, R.G.	258
alpha, coefficient	160, 163, 165t, 167, 168, 169, 170, 171, 194–95
alpha level (α)	227, 243, 253
<i>see also</i> confidence level	
American Council for the Teaching of Foreign Languages (ACTFL) oral proficiency guidelines	301
American Educational Research Association	154, 259
American Psychological Association	154, 259
analysis of variance (ANOVA)	
assumptions for independence	245, 249

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

Index 345

- populations, normal distribution
 - of 245, 248–49
- robustness, of assumption
 - violations 245
- between-group variance (S_B^2) 244
- calculations
 - between-groups mean square (MS_B) 246, 249
 - between-groups variance 246
 - F-ratio 245, 247
 - sum of squares (SS_B), between-groups 245–46, 249
 - sum of squares (SS_W), within-groups 246, 249
 - within-groups mean square (MS_W) 247, 249
 - within-groups variance 247
- dependent variable method 244
- differential design (non-equivalent groups) 291–92
- groups, selection of 243–44
- illustrative example
 - 1-state statistical hypothesis and specify confidence level 248
 - 2-check assumptions 248–49
 - 3-calculate mean squares 249
 - 4-calculate F-ratio 249
 - 5-determine F-ratio level of significance and interpret results 249–50
- descriptive statistics, source table for 248
- independent variable method 244
- multivariate 255
- planned comparisons 250–51, 256
- post hoc comparisons 250–52, 256
- repeated measures ANOVAs 255
- t-test, as inappropriate for more than two groups 243
- total variance (S_T^2) 244
- two-way 255
- within-group variance (S_W^2) 244, 247
- Anderson, J.C. ix–xii
- Anderson, N. 276–77
- ANOVA *see* analysis of variance
- applied linguistics ix–x, 4
- area conversion 313
- Asher, H.B. 111
- assessment
 - assessment/measurement/test, relationship among 9
 - meanings of 6–7
 - replicability 7
 - results, qualitative and quantitative reporting of 8
 - settings for 6–9
 - substantive requirement 7
 - systematicity requirement 7
- assumptions, of statistical tests
 - analysis of variance 245, 248–49
 - correlation coefficients 99
 - Pearson product-moment correlation coefficient (r) 85, 90
 - reliability estimates 163, 165–66
 - Spearman rank correlation coefficient (r_s) 88
 - t-test 236, 238, 240
- asymmetry 50, 61
- attenuation 93–94
- attributes, measurement of
 - unobservable 8
- averages
 - arithmetic 56
 - compensatory composite scores 318
 - indicators of central tendency 75
- axes 46
- B**
- Bachman, L.F. x, 4, 5, 8, 9, 14, 15, 17, 23, 27, 32, 118, 120, 122, 137, 146, 147, 149, 151, 153, 155, 156, 165t, 183, 184–85, 233, 247, 264, 269, 272–75, 281, 283–87, 301
- band descriptors
 - differences in 296–97, 298

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

346 INDEX

- band descriptors (*cont.*)
 - levels of ability 301
 - ordered domains 299
- bar charts 46, 47
- Berk, R.A. 202
- Berry, W.D. 111
- beta level (β) 227, 243
- between-group design (multiple-group/independent/uncorrelated) 103, 132–33, 217–18, 228
- between-group variance (S_b^2) 244
- 'bias' analysis 149
- bimodal distribution 50, 53
- biserial correlation coefficient (r_{bis}) 92, 101, 129
- bivariate distribution 33, 78–79, 100
- Bonferroni post hoc comparison 251, 252
 - see also* Analysis of Variance, post hoc comparisons
- Borich, G. 64
- box-and-whisker plots 47, 48
- Brennan, R.L. 183, 194
- Brown, J.D. 30, 32, 118, 182, 183, 194, 202, 268, 291–92
- Buck, G. 4
- C**
 - Campbell, D.T. 282
 - Canale, M. 14, 283
 - Carroll, J.B. 8–9, 14, 91, 279
 - Castellan, N.J. 90
 - CAT *see* computer adaptive testing
 - categorical variables *see* variables, discrete
 - ceiling effect 96
 - central tendency *see* grouping
 - certification decisions 11
 - CFA *see* confirmatory factor analysis
 - Chapelle, C. 14
 - chi-square statistic 277
 - Choi, I.-C. 149, 233, 247, 272–75, 281
 - CI *see* confidence intervals
 - CLA *see* communicative language ability
 - Clapham, C. 9, 15, 17, 93, 118, 137, 301
 - classical item analysis (IA)
 - dichotomous scoring scales 122
 - distractors 122, 130
 - item banks 135, 139–40
 - item difficulty 122, 125–27, 131–32
 - item discrimination 92, 122–23, 127–28, 132–33, 134, 135, 144, 151
 - item Record Forms 135–36
 - item selection 137–38
 - item-total score correlations 129
 - limitations of 139–41, 152
 - partial credit scoring scales (P-C) 122
 - procedure for 131
 - purposes of 121
 - response tallies 123–25
 - test improvement 135–36
 - test mean, calculation of 129–30
 - test variance 130
 - classical test theory (CTT)
 - equation expressing 157
 - measurement error
 - assumption of uniformity across levels 174, 175
 - multiple sources of 160, 171, 174–75
 - systematic *versus* random 157, 174, 175
 - precision of measurement 189–90
 - relative and absolute error 187
 - reliability estimates
 - correlational approach 160
 - inconsistency over time (stability) 166–68, 171
 - independence and parallel measurement assumptions 163, 165–66

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

Index 347

- internal consistency 161–66, 195
- rater consistency 169–70
- test form inconsistency 167–68
- variance approach 160 (*see also*
 - coefficient alpha; Kuder-Richardsons estimates)
- sources of inconsistency, combined effect of 175
- ‘true score variance’ 157–58, 191
- classification errors *see* dependability
- codes of practice 35, 154, 259
- coefficient alpha 160, 163, 165t, 167, 168, 169, 170, 171, 194–95
- coefficient kappa (κ) 200–202
- coefficient of determination 103–4
- coefficient \hat{p}_o (estimated proportion of agreement) 200, 202
- Cohen, A. 258, 279
- communicative language ability (CLA) 273–74
- composite scores 295
 - compensatory or non-compensatory 318–19, 322
- composites *versus* profiles 317–18
- different abilities, measures of 317
- intended interpretations and uses 316–21
- same ability, measures of 316–17
- underlying constructs, need for clear definition of 316, 317, 322
- weighting
 - effective weight 320–21
 - nominal weight 319–20
 - self-weight 320
- computer adaptive testing (CAT) 150
- confidence intervals (CI)
 - accuracy/confidence trade-off 235
 - calculation of 172–73
 - independence 233
 - interval estimate 231–32
 - large-sample means, differences between 232–35
 - normal distributions 233–34
 - point estimates 231
 - probability levels and ranges 173–74
 - z-scores
 - versus* t-test 242
 - use of 232, 233, 235
- confidence level 227, 232–35, 237–38, 240, 241, 248
 - see also* errors, in hypotheses testing; level of significance
- confirmatory factor analysis (CFA) 113
 - factor loadings, hypothesized 283, 284–85
 - illustrative study 283–86
 - model-data fit 283, 285–86
 - validation argument 286–87
- construct definition
 - conceptual 14–15
 - operational 15–16, 18–19, 27
- construct validity
 - demonstrating evidence for 5–6
 - measurement scales 27
 - representativeness 28
 - score interpretations 19
 - sources of variance 155–56
- constructs, and variables 18–19
- contingency coefficient (C) 92
- continuous variables 24–25, 26t, 46, 47, 129
- continuum 31
- control group 289
- correlated group design (single-group/within-group/dependent) 217, 228
- correlation
 - definition of 80–81
 - pre-existing expectations 79, 83–84, 108
- correlation coefficients
 - appropriate use of 84–85
 - causality 108–9
 - covariance 80, 103–4

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

348 INDEX

- correlation coefficients (*cont.*)
 - dependent and independent variables 105
 - different types 91–92
 - distributional characteristics
 - assumptions, checking 99
 - bivariate distribution, checking 100
 - combining multiple groups 98–99, 103
 - curvilinear relationships 95, 100
 - extreme cases 97–98, 101–2
 - range restrictions 96–97, 101
 - unequal variation (heteroscedasticity) 97, 100
 - univariate distribution, 100
 - error of prediction 107
 - meaningfulness of 253
 - measurement error
 - attenuation 93–94
 - reliability coefficients, calculation of 93–94
 - Pearson and Spearman, comparison between
 - non-normal distributions 100
 - strength and direction, interpretation of 89–90, 91
 - test scores, distributional characteristics of (skewness and kurtosis) 90–91
 - tied ranks, effect of 90
- Pearson product-moment
 - correlation coefficient (r) 85–87, 90, 92t, 253, 342
- regression equation 105–7
- regression intercept 106–7
- regression line 105–6
- regression weight 106–7
- relationships, characteristics of 81–84
- single correlation coefficient, inherent ambiguity in interpretation of 108–9
- Spearman rank correlation
 - coefficient (r_s) 87–89, 92t, 253, 343
 - statistical significance of 251, 253
 - as tool for test development 79
 - variables, strength and direction of relationship between 79, 82, 84–85, 89–90, 91, 113–14
- correlational procedures
 - factor analysis 111–12
 - multiple linear regression analysis 110–11
 - path analysis 111
 - structural equation modeling (covariance structure analysis) 112–13
- correlations, real or chance 210–11, 214
- costs, of decision errors 11, 12–13, 199
- counterbalanced design 168
- covariance 80, 103–4
- covariance structure analysis 112–13
- CR tests *see* criterion-referenced tests
- Crick, J.E. 183
- criterion level (mastery level) 198
- criterion-referenced scores
 - absolute decisions, appropriateness for 294, 300, 321
 - description of tasks 300
 - domain specificity 299, 300
 - levels of ability 301
 - number of tasks 300
 - ordered/unordered domains 299
 - percentage correct score 300–301
 - test contents, differences in 298–99
- criterion-referenced tests 30, 31–32
- Crooks, T. 258
- CTT *see* classical test theory
- cumulative frequency 307
- curvilinear relationships 95, 100
- cut score 198, 199, 204–5

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

Index 349

DD-study *see* decision study

data collection design 216

calculations, use of appropriate 218

multiple-group design 217–18, 228

single-group design 217, 228

data handling procedures 102

Davidson, F. 15, 32, 102, 118, 149, 233,
247, 272–75, 275, 281

Davis, J.A. 111

decision errors 11, 12–13, 199

decision study (D-study)

absolute error variance 187–88

generalizability coefficient (ρ_2) 188index of dependability (Φ) 188–89observed scores, dependability of
184

optimization 188

procedure 183

purpose of 183

random *versus* fixed facets 183–84

relative error variance 186–87

total score variance 186

universes 183–84

variance components,
interpretation of 184–85

decisions

absolute *versus* relative 10–11

relative importance of 11–13

see also absolute decisions; relative
decisions

degrees of freedom 250, 253

dependability

absolute criteria 192

agreement indices 198–204, 205

classification errors

false negative 199

false positive 199

misclassifications 199

cut score 198, 199, 204–5

domain scores 193–94, 198

factors affecting

set cut score 204–5

test length 204

total score distribution, statistical
properties of 204generalizability theory 178–79, 186,
191index of dependability (Φ) 188–89,
194–95, 205

mastery level 198

non-normal distribution 193

norm-referenced tests 192–93

standard error of measurement,
and confidence intervals
195–97

test score classifications 193–94

dependent group design (single-
group/within-group/
correlated) 217, 228dependent variable 105, 110–13, 244,
255

descriptive statistics 32, 33

test development and use,
applications to 75–76, 113deviation scores 58–59, 66–68, 85–86,
310–11

dichotomous scale 20

dichotomous variables

and continuous variables,
correlation between (point
biserial) 129

mean of 57, 125

range restriction 101

scoring scales (Right-Wrong) 20,
122differential design (non-equivalent
groups)ability intended to be
measured/ability basis of
grouping, mismatch between
292

analysis of variance, use of 291–92

illustrative study 291–92

implementation, ease of 292

a priori grouping 291, 292

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

350 INDEX

differential design (non-equivalent groups) (*cont.*)
 same test, administration of to all groups 291
 validation argument claim 291
 directional statistical hypotheses 215–16, 240
 dispersion *see* variability
 distractors 122, 126, 130
 distribution-free statistics 254
 distributions *see* score distributions
 domain-referenced tests 32, 299
 domain scores 193–94, 198
 domain scores (analogue of CTT ‘true score’) 193–94, 198
 Douglas, D. 4
 Draper, N.R. 111

E

Ebel, R.L. 130
 ecological correlation 98–99, 103
 Educational Testing Service (ETS) 259
 EFA *see* exploratory factor analysis
 effect size 254
 effective weight 320–21
 Embretson, S.E. 146, 151
 equal variance (homoscedasticity) 236–38–239, 245
 equivalence (equivalent forms reliability) 167–68
 Ericson, K.A. 279
 error variance
 absolute 187–88
 relative 186–87
 errors, in hypotheses testing 227
 eta (η) coefficient 92t
 eta² (η^2) coefficient 92t
 ETS *see* Educational Testing Service
 evaluation x, 5, 9
 evidence-centred design 263–64
 Excel (Microsoft 2002) 46
 experimental design (equivalent groups)

components of
 different treatments 288
 post-test 288
 random assignment to groups 288
 random selection 287, 288
 conditions to be satisfied
 different treatments 290
 effective instruction 290
 randomization 290
 control group 289
 experimental group 289
 illustrative study 288–90
 exploratory factor analysis (EFA)
 example study 279–82
 factor loadings, interpretation of 280–83
 extreme cases 101–2
 facilitating cases 97–98, 102
 outliers 49, 97–98, 102

F

F-ratio 159, 236, 238–39, 245, 247, 249–50
 F-values table 337–41
 facets of measurement 146, 176–78, 179, 180, 182, 183–84
see also generalizability theory
 FACETS program 147–49
 facilitating cases 97–98, 102
 factor analysis 111–12
see also confirmatory factor analysis; exploratory factor analysis
 false positive/negative classification errors 199
see also dependability
 fatigue, test-taker 287, 292
 Feldman, S. 111
 Fink, A. 34
 first quartile 64
 Fisher’s Z-transformation 170
 Fiske, D.W. 282

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

Index 351

fixed facets 183–84
see also generalizability theory
 flat distribution *see* platykurtic
 distribution
 floor effect 97
 frequency distribution
 graphical displays of 52–54
 of grouped scores 43–45, 48, 49
 interval scores 44–45, 47, 48
see also grouping; score
 distributions
 frequency polygons 46–47, 48, 50–53
 Fruchter, B. 70, 129

G

G-study *see* generalizability study
 G-theory *see* generalizability theory
 Gamma coefficient 92t
 generalizability coefficient (ρ_2), for
 relative decisions 188
 generalizability study (G-study)
 item difficulty, and mean of facet
 condition 180
 sources of variance 179–80, 181, 182
 total score variance 180, 181, 182
 generalizability theory
 and classical test theory 189
 dependability 178–79, 191
 designs
 single-facet crossed 179–80
 two-facet fully crossed 180–81
 two-facet nested 181–82
 facets of measurement 176–78, 179,
 180, 182
 fixed facet 183–84
 random facet 183–84
 multiple sources of variance,
 measurement of 176
 precision of measurement 189–90
 test scores, generalizability of 263
 universe score 178
 universes
 and D-study 183–84

universe of admissible measures
 176–77
 universe of generalization 177–78
see also decision study;
 generalizability study
 GENOVA (G-theory software) 183
 Glaser, R.W. 32
 Glass, G.V. 38, 44, 46, 74, 90, 91, 92,
 100, 170, 236, 245, 251, 254
 Gorsuch, R.L. 112
 grading on the curve
 norm-referenced tests 30
 relative decisions 11
 graphical displays of data
 bar charts 46, 47
 box-and-whisker plots 47, 48
 frequency distributions 52–54
 frequency polygons 46–47, 48,
 50–53
 histograms 46, 47
 normal distribution 45–46
 scatterplots 81–82, 100
 Green, E. 278, 279
 Gronlund, N.E. 4, 259, 315
 grouping (central tendency)
 appropriate uses of indicators of
 61–63
 frequency distributions/graphical
 displays, limitations of 52–54
 levels of measurement 54
 mean 56–60, 63
 median 55, 62–63, 73
 mode 55, 61, 62, 73
 mode/median/mean, relationships
 among 61, 62
 reporting of 73–74
 Guilford, J.P. 70, 91, 129, 320
 Guttman split-half reliability estimate
 162, 165t, 171

H

Hambleton, R.K. 142, 144, 145, 151,
 175

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

352 INDEX

- Harmon, H.H. 112
- HAX/VOY axis mnemonics 46
- heterotrait correlations 282
- heteroscedasticity 97, 100
- Hinofotis, F.B. 279
- histograms 46, 47
- homoscedasticity 236, 238–39, 245
- honestly significant difference (HSD) 251, 252
- Hoover, H.D. 320
- Hopkins, J.D. 38, 74, 91, 92, 100, 236, 245, 251, 254
- Hopkins, K.D. 4, 30, 44, 46
- horizontal axis (abscissa) 46
- Hudson, T. 30, 32, 118, 182, 183, 202
- hypotheses
 - beliefs/hunches/questions 212, 213
 - confidence levels, specifying 227
 - directional statistical hypotheses 215–16, 240
 - non-directional statistical hypotheses 215, 216, 233, 238
 - null hypothesis 214, 216, 226, 227, 228, 233, 243
 - operational research hypotheses 212, 213, 228
 - statistical hypotheses, appropriate choice of 216
 - statistical research hypothesis 212, 213, 214–16, 228
 - testing, levels of significance and errors in 227
 - theoretical research hypothesis 212, 213, 228
- I**
- IA *see* classical item analysis (IA)
- ICC *see* item characteristic curve
- ICF *see* item characteristic function
- IFF *see* item information function
- independence, of samples 237, 238, 240
- independent group design (multiple-group/between-group/uncorrelated) 103, 132–33, 217–18, 228
- independent variable 105, 110–13, 244, 255
- inferential statistics 34–35
- inter-rater reliability 169
- International English Language Testing System* (IELTS) band descriptors 301
- International Language Testing Association 154
- interpretive arguments (validation)
 - claims and counterclaims, evidence in support of 264–67, 269, 270–71, 293
 - demonstrating plausibility of 262
 - evidence-centred design 263–64
 - evidentiary reasoning 258
 - inferences 267–68, 269
 - links in 262–63
 - nature of 262
 - observation/observed score/target score links 262–63
 - score use, consequences of 261–62 *see also* validity
- interval estimate 231–32
- interval scales
 - amount of difference 21–22
 - continuous variables 24–25
 - exhaustive 44
 - frequency distribution 44–45, 47, 48
 - mutually exclusive 44
 - properties of 26
 - standard deviation (S) 65–66
- intra-rater reliability 169
- IRT *see* item response theory
- item analysis *see* classical item analysis (IA)
- item banks 135, 139–40
- item characteristic curve (ICC) 142–44
- item characteristic function (ICF) 142–44

- item difficulty 122, 125–27, 131–32, 140, 151
 - item discrimination
 - CR item discrimination index (D_{m-nm}) 132, 135
 - definition of 144, 151
 - interpretation of 130–31
 - multiple-group design 132–33
 - norm-referenced tests 122–23, 127–28
 - point-biserial correlation coefficient (r_{pbis}) 92
 - pre-post-test difference index (DIS_{ppd}) 132–33, 134, 135
 - R-W and P-C scored items 127–28, 132
 - single-group design 132
 - test variance-discrimination index relationship 130
 - item information function (IIF) 144–46, 150, 190
 - item response theory (IRT)
 - advantages of 142
 - calculations, difficulty of 151
 - classical item analysis, advantages over 151
 - computer adaptive testing 150
 - item characteristic curves 141, 142–44
 - item information function 144–46, 150, 190
 - item parameters
 - 1-parameter model (Rasch model) 141, 152
 - 2-parameter model 141, 152
 - 3-parameter model 141, 152
 - a-parameter (discrimination) 141, 144, 145
 - b-parameter (IRT difficulty) 141, 143–44, 145
 - c-parameter (guessing) 141, 143, 145
 - model-data fit, importance of 142
 - precision of measurement 189–90
 - test information function (TIF) 150, 190
 - underlying traits, assumption of 141
 - unidimensionality 141
 - item specifications 135, 137
 - item variances
 - coefficient alpha, use of 160, 163, 165t, 171
 - data collection approach 163
 - Kuder-Richardsons estimates 163–64, 165t, 171
 - longer test, reliability of 164–65
- K**
- Kane, M. 258, 262–63, 264
 - kappa coefficient ($\hat{\kappa}$) 200–202
 - kappa squared ($\hat{\kappa}^2_{(X,T_X)}$) 203–4
 - Kendall tau (τ)² coefficient 92t
 - Kim, J.-O. 112
 - Klaus, D.J. 32
 - Kolen, M.J. 320
 - Kubiszyn, T. 64
 - Kuder-Richardsons estimates 163–64, 165t, 171, 194
 - Kunnan, A.J. 113, 183
 - kurtosis 50, 74–75, 90–91, 100, 233–34
 - see also* leptokurtic distribution;
 - mesokurtic distribution;
 - platykurtic distribution; score distributions
- L**
- Lado, R. 14
 - language assessment
 - applied linguistics context 4
 - evaluation, as use of 9
 - measurement context 4
 - large-scale standardized tests, norming of 304
 - large-scale testing agencies 30

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

354 INDEX

- leptokurtic distribution (peaked) 50, 51
see also kurtosis; mesokurtic distribution; platykurtic distribution
- level of significance
 F-ratio 249–50
 hypothesis testing 227
 t-ratio 237, 239, 241–42
see also confidence level
- levels of measurement *see* measurement scales
- Lewis-Beck, M.S. 111
- Lickert scale 17
- Linacre, J.M. 147
- linear relationships 82
see also non-linear relationships
- linear transformation 60, 72
 linear t-scores 312–13, 313
 z-scores 310–12, 313
- Linn, R.L. 4, 32, 259, 300, 306, 315
- logit scale 147, 148t
- Lomax, R.G. 113
- Lowe, P. 14
- Lukmani, Y. 275
- Lumley, T. 275
- Luoma, S. 4
- Lynch, B.K. 15, 32, 118, 147, 184–85
- M**
- MANOVA *see* multivariate analysis of variance
- many-facet Rasch measurement (MFRM)
 facets of measurement 146
 model-data fit 146, 147, 148t
 relative difficulty 146, 147, 148
 relative inconsistency 146–47, 148–49
 relative severity 147, 148
- Marcoulides, G.A. 113
- Mason, M. 147, 184–85
- mastery level 198
- McNamara, T.F. 4, 146, 151, 275
- mean
 calculation of 56–57
 definition of 56
 dichotomous variables 57, 125
 extreme scores, effect of 58–60
 as fulcrum 58–59
 item selection 137
 mode/median/mean, relationships among 61, 62
 raw score interpretation 302–4
 sample mean and population mean, relationship between 219, 220–21
 scaled item mean (p^*) 126–27, 129–30
 transforming by constant 60
 uses of 75
- measurement
 definition of 8
 limitations on
 imprecision 29
 incompleteness 28
 indirectness 28
 relativity 29–30
 subjectivity 29
 underspecification 27–28
- measurement error
 assumption of uniformity across levels 174, 175
 correlation coefficients 93–94
 D-studies 186–88
 multiple sources of 174–75
 reliability 93–94, 153
 sources of 160, 171
 systematic *versus* random 157, 174, 175
see also standard error of measurement
- measurement process, steps in
 define construct conceptually 14–15
 define construct operationally 15–16, 18–19, 27
 elicited responses

- rating scales (judging approach) 16–18
 - scoring individual tasks (counting approach) 16, 17
 - as essential for statistical analyses 19
 - observational data 16, 17
 - attribute categories, assigning numbers to 17
 - attribute occurrences, counts of 17–18
 - quantify observations 16–18
 - reliability, demonstration of 19
 - variables, and constructs 18–19
 - measurement properties 18
 - measurement scales 19
 - construct validity 27
 - dichotomous scales 20
 - grouping indicators 63
 - interval scales 21–22, 23, 24–25, 26
 - nominal scales 20, 24, 26
 - ordinal scales (ranking) 20–21, 24–25, 26
 - ratio scales 22–26
 - relationships among 26–27
 - reliability 27
 - variables, discrete *versus* continuous 24–25
 - median
 - box-and-whisker plots 47, 48
 - calculation and definition of 55–56
 - mode/median/mean, relationships among 61, 62
 - ordinal scales 62–63, 73
 - mesokurtic distribution (normal) 50, 51, 309
 - see also* kurtosis; leptokurtic distribution; platykurtic distribution; score distributions
 - Messick, S. 258, 259, 276
 - MFRM *see* many-facet Rasch measurement
 - mGENOVA (G-theory software) 183
 - Milanovic, M. 275
 - Mislevy, R.J. 257, 258, 263
 - mode
 - definition of 50, 51–52, 54
 - item selection 137–38
 - mode/median/mean, relationships among 61, 62
 - nominal scales 61
 - ordinal scales 61, 62, 73
 - uses of 75
 - model-data fit 142, 147, 148t, 283, 285–86
 - see also* confirmatory factor analysis; many-facet Rasch measurement
 - monomethod correlations 282
 - monotrait correlations 282
 - MTMM *see* multitrait-multimethod correlation matrix
 - Mueller, C.W. 112
 - multi-point scale 17
 - multiple-group design (between-group/independent/uncorrelated) 103, 132–33, 217–18, 228
 - multiple linear regression analysis 110–11
 - multitrait-multimethod correlation matrix (MTMM) 282–83
 - conditions to be satisfied 287
 - example study 283–87
 - ordering test taking 287
 - test taker fatigue 287
 - multivariate analysis of variance (MANOVA) 255
 - multivariate distribution 33
- N**
- N-way analysis of variance 255
 - National Council on Measurement in Education 154, 259
 - negatively skewed distribution 50, 53, 61, 62
 - see also* score distributions

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

356 INDEX

- Nevo, N. 276
 - Nitko, A.J. 304, 306, 313
 - nominal scales
 - dichotomous scales 20
 - discrete variables 24
 - distinctiveness 20
 - mode 61
 - properties of 26
 - variability 72, 73
 - nominal variables 91–92
 - nominal weight 319–20
 - non-directional statistical hypotheses 215, 216, 233, 238
 - non-equivalent groups design *see* differential design
 - non-linear relationships 95, 96–97, 100, 101
 - non-linear transformation 100
 - non-parametric tests of significance 254
 - norm groups 30, 304
 - norm-referenced (NR) tests 29–30, 32
 - dependability 192–93
 - norm-referenced scores
 - comparability 315
 - comparison of types of 313–14
 - level and form specific 308
 - ordinal scale 308–9
 - percentile ranks 306–9, 310, 313, 314
 - rankings 306
 - reference groups, norming 304, 305, 308, 315
 - relative decisions 294, 322
 - relative ease/difficulty of test, irrelevance of 305
 - relevance 315
 - representativeness 315
 - score variation, and reliability 305
 - standard scores 310–14
 - types of 306–13
 - up to date 315
 - uses of 304–5
 - normal distribution
 - as ‘chance’ 222–24
 - characteristics of 221–22, 309–10
 - confidence intervals 233–34
 - graphical representation 45–46
 - non-normal distributions 100
 - norm-referenced tests 30–31
 - null hypothesis, testing of 226
 - percentile ranks 310, 314
 - Norris, J.M. 268
 - North, B. 301
 - NR tests *see* norm-referenced (NR) tests
 - null hypothesis 214, 216, 226, 227, 228, 233, 243
- O**
- object of interest 7
 - Oller, J.W. 14, 279, 282
 - one- and two-tailed tests
 - appropriate uses of 226
 - directional research hypotheses 225–26
 - non-directional research hypotheses 224–26
 - null hypothesis, appropriate test for rejection of 226
 - one-way analysis of variance *see* analysis of variance
 - operational research hypotheses 212, 213, 228
 - optimization 188
 - ordinal scales
 - continuous variables 24–25
 - correlation coefficients 87, 91–92, 92t
 - distinctiveness and ordering 21
 - median 62–63, 73
 - mode 61, 62, 73
 - percentile ranks 308–9
 - properties of 26
 - semi-interquartile range 64, 73
 - use of 308–9
 - ordinate (vertical) axis 46

outliers 49, 97–98, 102

see also correlation coefficients

P

Palmer, A.S. x, 9, 15, 17, 120, 122, 137, 153, 156, 264, 269, 283–87, 301

parallel measures 159, 161, 162

parametric tests 231

path analysis 111

peaked distribution *see* leptokurtic distribution

Pearson, K. 85

Pearson product-moment correlation coefficient (r) 253, 342

assumptions to be met 85, 90

calculation of 85–87

interval data 92t

and Spearman rank correlation coefficient (r_s) 89–91, 100

Pedhazur, E.J. 111

percentage correct score 300–301

percentile ranks

advantages of 308

calculation of 306–8

and normal distribution 310, 314

versus percentage scores 308

relative standing of test taker 308

sample size, not dependent on 308

and standard scores, comparison between 313–14

Peterson, N.S. 320

Phakiti, A. 276

phi coefficient (Φ) 92, 188–89, 194–95, 205

phi lambda (Φ_λ) 203–4

planned (*a priori*) comparisons 250–51, 256

platykurtic distribution (flat) 50, 52

see also kurtosis; leptokurtic distribution; mesokurtic distribution; score distributions

point-biserial correlation coefficient (r_{pbis}) 92, 101, 129

point estimates 231, 242

Pollitt, A. 17

Popham, W.J. 30, 31

population

definition of 34

normal distribution of 236, 238, 245, 248–49

parameters 35, 36, 220

sample and population mean, relationship between 219, 220–21, 223–24

see also unbiased estimator of the population standard deviation (s^2)

population correlation coefficient (ρ) 88, 92t

positively skewed distribution 50, 52, 61, 62

see also score distributions

post hoc (*a posteriori*) comparisons 250–53, 256

Bonferroni 251, 252

Tukey's honestly significant difference (HSD) 251, 252

pre-post-test difference index (DIS_{PPD}) 132–33, 134, 135

precision of measurement 189–90 *see also* classical test theory

prediction

correlation as a regression line 105–7

multiple linear regression analysis 110

path analysis 111

pretesting 119–20

probabilities, one- and two-tailed 224–26

Purpura, J.E. 4

Q

qualitative data 6, 16

quantitative data

activities and procedures for collection of 6

358 INDEX

- quantitative data (*cont.*)
 appropriate analysis 13
 as evidence in support of test use 3
 statistical analysis 3–4
 see also measurement process;
 measurement scales
- quartiles 64–65
- R**
- random facets 183–84
 see also facets of measurement;
 generalizability theory
- random measurement error 157, 174,
 175
- randomization 290
- range 50, 51–52, 54, 63–65, 69–70, 75
- range restriction 96–97, 101
- ranked data, ranking *see* ordinal scales
- Rasch model 141, 146, 152
 see also many-facet Rasch
 measurement
- rater consistency
 coefficient alpha, use of 169, 170
 data collection design 169
 inter-rater reliability 169
 intra-rater reliability 169
 parallel measures assumption,
 violation of 169–70
 ratings, correlation between 169–70
 see also classical test theory
- rating scales 16, 17, 28–29, 273–74,
 298, 301
- ratio scales 23–24
 continuous variables 24–25
 properties of 26
- raw scores
 band descriptors, differences in
 296–97, 298
 common scale, need for 297, 303
 deviation scores 66
 mean 302–4
 meaningful interpretation of 294,
 298, 321
- percentile ranks *versus* standard
 scores 313–14
- problems interpreting 302–3
- scales of measurement, differences
 in 297–98, 301–2
- score distributions 42
- standard deviation 68–69, 302–4
- test differences 296–97, 302–3
- test-takers, relative standing of
 302–3
 see also confidence intervals
- Raykov, T. 113
- Read, J. 4
- reference groups, norming 304, 305,
 308, 315
 see also norm-referenced scores
- regression
 equation 105–7
 intercept 106–7
 line 104–6
 weight 106–7
 see also multiple linear regression
 analysis
- Reise, S.P. 146
- relative decisions 294, 322
 versus absolute decisions 10–11, 192
- relative error variance 186–87
- reliability
 equivalence 167–68
 individual test scores 171–74
 longer tests 164–65
 measurement error 93–94, 153–54
 operational definition of 158–59
 independence, meeting
 assumption of 159, 161, 162
 parallel measures, conditions of
 159, 161, 162
 professional responsibility for 154
 relative *versus* absolute decisions
 192
 standard error of measurement,
 calculation of 172, 191
 test length 190

- theoretical definition of 153, 157–58
- total score distribution, statistical
 - properties of 190
 - see also* classical test theory; confidence intervals; generalizability theory; item response theory; variance, sources of
- replicability 7
- reporting scores *see* scores, reporting and interpreting
- rho coefficient (r_s) 88, 92t
- robustness 38, 236, 245
- Rogers, H.J. 142, 144, 145, 151, 175
- rounding, of decimals 57n, 67, 86, 89
- Ryan, K. 149, 233, 247, 272–75, 281
- S**
- sample independence 237, 238, 240
- sample statistics 210–12, 230–31
 - accuracy, of estimates 35–36
 - population parameters 35, 36, 220
 - sample representativeness 36, 219
- sampling distributions
 - means, comparison of 220–22
 - sample and population mean, relationship between 219, 220–21, 223–24
- sampling error 36, 63, 70–71, 211
- Sasaki, M. 276
- scaled item mean (p^*) 126–27, 129–30
- scales *see* measurement scales
- scatterplots 81–82, 100
- Schumacker, R.E. 113
- score distributions
 - descriptive statistics, purposes of calculating 42
 - frequency distribution
 - of grouped scores 43–44, 48, 49
 - of interval scores 44–45, 47, 48
 - graphical representation of
 - box-and-whisker plots 47, 48
 - frequency polygons 46–47, 48, 50–53
 - histograms (bar charts) 46, 47
 - negatively skewed distribution 50, 53, 61, 62
 - ordered listing of scores 43
 - positively skewed distribution 50, 52, 61, 62
- raw scores 42
- shapes of
 - asymmetricality 50, 61
 - bimodal distribution 50, 53
 - distributions, differing
 - characteristics of 51–52
 - kurtosis 50, 74–75, 90–91, 100, 233–34
 - skewness 50, 52–53, 61, 62–63, 64, 74, 233–34
 - symmetricality 50, 61
- variable, distribution for 42
- see also* grouping
- scores, reporting and interpreting
 - test developer perspective 294
 - test user groups, reporting scores meaningfully for 294, 295–96
 - see also* composite scores; criterion-referenced scores; norm-referenced scores
- second quartile 64
- self-assessment 13
- self-weight 320
- SEM *see* standard error of measurement
- semi-interquartile range 64–65, 75
- Shavelson, R.J. 38, 80, 92, 183, 254, 255
- Siegel, S. 90, 254
- significance level *see* level of significance
- Simon, H. 279
- single-group design (within-group/dependent/correlated) 217, 228
- skewness
 - calculating 74

360 INDEX

- skewness (*cont.*)
 confidence intervals 233–34
 negatively skewed distribution 50, 53, 61, 62
 positively skewed distribution 50, 52, 61, 62
 score distribution shapes 50, 61, 62–63, 64, 90–91
see also score distributions
- small sample test *see* t-test
- Smith, H. 111
- Snedecor, G.W. 70
- Somer's D coefficient 92t
- Spearman-Brown prophecy formula 164
- Spearman-Brown split-half reliability estimate 161–62, 164, 165t
- Spearman rank correlation coefficient (r_s)
 assumptions to be met 88
 calculation of 88–89
 interval data 92t
 ordinal data 87, 92t
 and Pearson product-moment correlation coefficient (r) 89–91, 100
- split-half reliability estimates
 Guttman 162, 165t, 171
 random *versus* rational approach 161
 Spearman-Brown 161–62, 164, 165t
- SPSS computer program xi, 46, 49, 71, 75, 129, 222–23
- squared-error loss agreement indices 203–5
- stability 171
 coefficient alpha, use of 167
 test administrations, time lapse between 167–68
 test-retest design, sources of inconsistency in 166–67
 test scores, correlation between 167
- standard deviation
 box-and-whisker plots 47
 calculating from deviation scores 66–68
 calculating from raw scores 68–69
 checking for reasonableness of 69–70
 degrees of freedom 71–72
 description of 65–66
 extreme cases 101–2
 interval scales 65–66, 73
 item selection 137
 Pearson product-moment correlation coefficient (r) 85–87
 raw score interpretation 302–4
 reasonableness, checking for 69–70
 sample statistics 35
 standard error of measurement 172
 standard error of the mean 223–24, 230–31
 t-scores 312–13
 transforming by constant 72
 usefulness of 75
 z-scores 310–12
see also unbiased estimator of the population standard deviation (s^2)
- standard error of estimation ($SE(\theta)$) 190
- standard error of measurement (SEM)
 absolute decisions 195, 196, 205
 and confidence intervals 195–97
 reliability 171–4, 191
- standard error of the mean 223–24, 226, 230–31
- standard scores
 linear T-scores 312–13
 normal curve, and standard deviation units 309–10
 normalized standard scores 313
 z-scores 310–12, 313
see also percentile ranks
- standards, professional 35, 154, 259

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)

Index 361

Standards for Educational and Psychological Testing 259
Standards for Quality and Fairness (ETS) 259
 statistical inference 218–19
 statistical procedures,
 appropriateness and
 meaningfulness of 37–38
 statistical research hypothesis 212,
 213, 214–16, 228
 Steinberg, L.S. 258
 structural equation modeling *see*
 covariance structure analysis
 substantive requirement, for
 assessment 7
 sum of squares 67, 245–46, 249
 Swain, M. 14, 283
 Swaminathan, H. 142, 144, 145, 151, 175
 systematic measurement error 157,
 174, 175
 systematicity requirement, for
 assessment 7

T

T-scores, linear 312–13
 t-test (small-sample test) 159
 correlated test, procedures for
 1-state statistical hypothesis and
 specify confidence level 240
 2-independent or dependent
 sample 240
 3-check assumptions 240
 4-calculate t-ratio 240, 241
 5-determine level of significance
 of t-ratio 241–42
 equal variance, of populations, use
 of F-ratio to check 236, 238–39
 as inappropriate for more than two
 groups 243
 independence of scores 236–37
 population for samples, normal
 distribution of 236, 238
 sample independence 237

score independence 236–37
 small sample, size of 235
 uncorrelated test, procedures for
 1-state hypotheses and specify
 confidence level 237, 238
 2-independent or dependent
 sample 238
 3-check assumptions 238
 4-calculate t-ratio 237, 239
 5-determine level of significance
 of t-ratio 237, 239
 violation of assumptions,
 robustness of t-test for 236
versus z-scores 242
 t-values table 237, 239, 241, 336
 target language use (TLU) 267–68,
 269, 272–75, 318–19
 task precision 29
 tau b coefficient 92t
 tau c coefficient 92t
 test, definition of 8–9
 test content analysis
 content representativeness
 272–75
 domain-limited inferences 275
 example study 272–73
 expert judgements, inconsistency of
 275
 language ability, aspects of
 272–73
 limitations and problems with
 approach 275
 rating scales
 communicative language ability
 (CLA) 273–74
 test method facets 273–74
 task characteristics 272–73
 test-taker performance 275
 TLU domains, difficulty defining
 275
 test design 117
 test form inconsistency *see*
 equivalence

362 INDEX

test information function (TIF) 150, 190
see also item response theory

test method facets (TMF) 273–74

Test of English as a Foreign Language, norming of 304

Test of English Writing (TEW) 280

test performance, underspecification 27–28

test-retest reliability *see* stability

test scores, describing (descriptive statistics) 41–77

test specifications 15

test-taking processes, analysis of performance 275–76
 verbal protocol analysis 276–79

test tasks, analysing 119–52

test usefulness
 construct validity 5–6
 definition of x , 5
 evidence, collection of 6
 pretesting 119–20
 professional responsibility 5–6
 reliability 5–6
 test performance, and evaluation of x , 5
 test users, responsibilities of 6

tetrachoric (r_{tet}) coefficient 92, 101

theoretical research hypothesis 212, 213, 228

third quartile 64

threshold loss agreement indices
 advantages and disadvantages of 202
 coefficient kappa ($\hat{\kappa}$) 200–202
 coefficient p_o (estimated proportion of agreement) 200, 202

tied ranks 90

TIF *see* test information function

TLU *see* target language use

TMF *see* test method facets

traits, measurement of 8

true scores
 domain scores 193–94, 198

universe scores 178

variance 157–58, 191
see also classical test theory

truncated samples 96–97, 101

Tukey's honestly significant difference (HSD) post hoc comparison 251, 252
see also analysis of variance

type I and II errors 227, 243

U

unbiased estimator of the population standard deviation (s^2)
 CTT estimates of reliability 160, 162, 163, 167, 168

F-ratio 239

parallel measures, and reliability estimates 159

Pearson product-moment correlation coefficient 85

point estimates 231

sampling errors 70–71

uncorrelated group design (multiple-group/between-group/independent) 103, 132–33, 217–18, 228
see also data collection design

unequal variation (heteroscedasticity) 97, 100

unimodal distribution 61

univariate distribution 33

universe score 178, 263

universes 176–78, 183–84
see also generalizability theory

urGENOVA (G-theory software) 183

usefulness *see* test usefulness

V

validation arguments *see* interpretive arguments

validation study, design of 265
 1-operational procedures, development of 270

Cambridge University Press

978-0-521-00328-5 - Statistical Analyses for Language Assessment

Lyle F. Bachman

Index

[More information](#)*Index* 363

- 2-state hypotheses about expected patterns of relationships or differences 270–71
 - 3-design study to collect relevant data 271
 - 4-collect data 271
 - 5-test hypotheses 271
 - validity
 - experimental design (equivalent groups) 287–90
 - exploratory factor analysis 279–82
 - multitrait-multimethod correlation matrix 282, 287
 - non-equivalent groups design 290–92
 - test content analysis 272–75
 - test-taking processes, analysis of 275–79
 - as theoretical concept, and quality of particular test use 257–58, 292–93
 - unitary concept
 - evaluative judgement 260
 - score use, consequences of 261–62
 - score interpretation 260–61
 - validity as matter of degree 259–60
 - variability
 - levels of measurement 54
 - nominal/ordinal/interval scales 72, 73
 - range 63–64, 72, 73
 - semi-interquartile range (Q) 64–65, 73
 - standard deviation (S) 65–72, 73
 - variance (S^2) 66
 - variables
 - bivariate distribution 33
 - construct operational definition 18–19, 27
 - and constructs 18–19
 - dependent and independent and MANOVA 255
 - multiple linear regression analysis 110–11
 - N-way ANOVA 255
 - path analysis 111
 - structural equation modeling 112–13
 - dichotomous and continuous 129
 - discrete (categorical) and continuous 24–25, 26t, 46, 47
 - factor analysis 111–12
 - interval 91–92
 - multivariate distribution 33
 - nature of relationship between 33
 - nominal 91–92
 - univariate distribution 33
 - variance components 179, 184–85
 - see also* generalizability theory
 - variance (S^2)
 - definition of 65–66
 - sources of
 - language ability 155–56
 - personal characteristics 155, 156
 - random factors 155, 156
 - simplified measurement models for 157
 - systematic *versus* unsystematic effects 156
 - test method 155, 156
 - test task characteristics 156
 - verbal protocol analysis
 - claims and counterclaims 277–78
 - verbal report evidence 276–79
 - vertical (ordinate) axis 46
- W**
- Wall, D. 9, 15, 17, 93, 118, 137, 301
 - Webb, N.M. 183
 - weighting, in composite scores 318
 - component score correlations/ variance 319, 320–21

364 INDEX

weighting, in composite scores (<i>cont.</i>)	Z
composite score calculation 321	z-scores
effective weights 319–21	calculating 311–12
factors determining 319	common scale, obtaining 311
nominal weight 319–20	deviation scores, calculating
Weigle, S.C. 4	310–11
within-group design (single/	effective weight 320–21
dependent/correlated) 217, 228	exact probabilities associated with
<i>see also</i> data collection design	222–23
within-group variance (S_w^2) 244, 247	as inconvenient and difficult to
Workbook xi–xii, 46, 49, 75	interpret 312
Wright, B.D. 147	statistical inference function
	312
X	tables for 330–35
‘x’ axis (horizontal, abscissa) 46	Z-transformation 170
	zero point, true 22–23, 297
Y	Zimmerman, D.W. 38
‘y’ axis (vertical, ordinate) 46	Zumbo, B.D. 38