absolute (criterion referenced) methods
  p-values, 152
  standard setting, 3, 52, 100–102, 103
  academic changes, 191–192, 196–197
Academy of Medical Royal Colleges, 71
acute care assessment tool (ACAT), 69
ageing population, 193–194
Angoff method, standard setting, 54, 101–102
anomalous pair detection, 198
anonymised marking, 169, 176
answer cueing, 20, 22, 23–25, 161
anxiety, 180
test (exam stress), 160, 162, 163
appeals, student, 171, 192
appraisal, 132–133
assessment, 1–11
  academic change, 191–192, 196–197
  assessable attributes, 5–8
  blueprinting, 3, 45, 105–106, 107, 110, 189
  competence, 2–3, 8–9, 38–56, 59
  computer-based, 106–111, 135–136, 137, 194–196
definition, 4
economic change, 192–193
evaluation contrasted, 4
evaluation of tests see psychometrics
examination boards, 168–173
examiners see examiners
feedback see feedback
future developments, 185–198
future developments see also above
knowledge, 6–8, 9, 16–36
need for, 4
performance, 2–3, 8–9, 38, 59–74
professionalism, 9–10, 76–92
psychometrics, 141–165
regulatory change, 187–191
research into, 191–192, 196–198
sociological change, 193–194
standard setting, 3, 52–54, 94–105, 152
student failure, 169–171, 176–182
teaching–learning–assessment
  alignment, 5, 45, 105–106, 107, 110, 189
technical developments, 194–196
trends, 2–3
validity see validity of tests
assessment centres, 195
assessors see examiners
attendance problems, students, 181
attitudes, professionalism, 90
behaviour (affective domain), 7, 8
behavioural assessments, 90, 91
bias
   examiner, 173
   in tests, 158–164
blind double marking, 169
Bloom’s taxonomy of assessable attributes, 5–8
blueprinting, 3, 45, 105–106, 107, 110, 189
Bologna Declaration, 169–170
borderline groups, sequential OSCE, 54–55
borderline methods, standard setting, 52–54, 101–102
borderline students, examiners’ options, 170
budget allocations, 192–193
candidate failure see student failure
career pathways, gender factors, 193
case-based discussions (CbDs), 68–69, 82–83
case specificity problem, 17, 18–19, 106, 154
Certificate of Completion of Training (CCT), 188–189
Certificate of Eligibility for GP Registration (CEGPR), 189
Certificate of Eligibility for Specialist Registration (CESR), 189
cheating, 163–164, 195–196, 197–198
checklist rating, OSCE scoring, 51, 174
classical test theory (CTT), 147–149, 154–155
clients see service user assessment;
   service user images; service user simulators
clinical practice assessments see competence assessment;
   performance assessment
clinical reasoning tests, 28–32
case-based discussion, 68
Cohen method, standard setting, 103–105
compensated marks, 170
competence assessment, 38–56
dimensions, 39
evaluation criteria, 39–40
long case examination, 39, 40
OSATS, 39, 44
OSCE see objective structured clinical examination
OSLER, 39, 40–42
performance and, 8–9, 38, 59
reliability, 153, 154, 175
test formats, 39
   see also specific formats
trends, 2–3
competitiveness, 180
complaints, student, 171
computer-based assessments, 106–111, 135–136, 137, 194–196
computer-based cheating detection, 164, 197–198
condoned fails, 170
certainty-based marking, 27–28, 161
consensus statements on performance assessment, 73
constructed answer questions, 16–17
   see also essay-type questions
construct validity, 146
   see also validity of tests
contexts, professionalism, 91
convergent validity, 158
core curriculum, 3
criteria-guided workplace-based assessment, 71, 72
criterion referenced (absolute) methods
   p-values, 152
   standard setting, 3, 52, 100–102, 103
   critical incidents, 84
Cronbach’s alpha, 151
cueing, 20, 22, 23–25, 161
cultural issues, 177–178, 193
curricula
regulatory change, 188, 189
teaching–learning–assessment
alignment, 45, 105–106, 107, 110, 189
cut score setting see standard setting
decision-making skills assessment
case-based discussion, 68
key feature problems, 28
script concordance items, 30
demographic change, 193–194
diagnosis script questionnaires, 30
script concordance items, 30–32
directly observed procedural skill
(DOPS), 67–68
discriminant validity, 158
dissertations, 32–35
divergent validity, 158
doctors’ interpersonal skills
questionnaire (DISQ), 132
double marking, 169
doves see examiners, leniency/severity
economic changes, 192–193
educational institutions
examination boards, 168–173
professionalism, 91–92
educational paradigm, 5
education budgets, 192–193
education research, 191–192, 197–198
emotions, feedback, 122
English language testing, 189
environmental factors, 91, 171
e-portfolios, 194–195
essay-type questions, 17–18
marking, 17–18, 19, 111
modified (MEQs), 17, 18–19
professionalism, 85
short answer (SAQs), 20, 85, 111
ethnicity, 177–178, 193
European Credit Transfer and
Accumulation System (ECTS), 169–170
European Economic Area,
qualifications, 188–189
evaluation
assessment contrasted, 4
of tests see psychometrics
exam failure see student failure
examination boards, 168–173
examination performance data,
196–197
examiners, 167–176
bias, 173
borderline student options, 170
designing assessment criteria, 168
error-reduction strategies, 177
external, 171–173
inter-rater reliability, 153–154, 162,
175, 176
judgement process, 173–175
leniency/severity, 10, 50, 173,
174–176, 177
marking see marking
moderation of assessments, 169, 172
OSCE circuits
feedback, 50
judgement process, 173–174
marking, 43, 47, 51–54, 82, 110,
174, 175
professionalism, 82
standard setting, 52–54
training, 48–50
of professionalism, 82, 83
verbal protocol analysis, 175–176
workplace-based assessments, 70–71,
174
exam marking see marking
exam questions see written questions
exam stress, 160, 162, 163
extended matching questions (EMQs),
17, 20, 22–23, 24, 25–26
external examiners (EEs), 171–173
failure at exams see student failure
feedback, 114–138
appraisal, 132–133
on assessment, 126, 134–136, 137
benefits, 116

case of the month scenarios, 135–136, 137

definitions, 115–116, 117

discussion, 122

facilitation challenges, 117–122, 124–126
facilitation models, 128–129
facilitators defined, 117
facilitator–learner relationship, 121–122, 125, 127–128
formative–summative balance, 119–120

teachers, 118, 119

future developments, 190–195
importance, 116–117
language factors, 120, 122
meta-feedback skills, 126, 134–136
multi-source, 9, 10, 87, 130–132
OSCE, 50
power relations, 122
practical suggestions, 126–129
principles, 124–126
range of sources, 117, 118
recipient factors, 120–122, 127–128

roles in facilitating, 117–122, 124–126

service user, 86–87, 130–132

theories, 123–124
typology, 118, 119
workplace-based assessment, 70–71, 190–191

formative assessments, 10, 95
appraisal, 133
balance with summative, 95, 119–120
feedback, 116, 119–120, 133
feedback on, 134–135

Kolb's learning cycle, 61
professionalism, 80, 83, 84

simulation, 60, 61–66, 108, 195
ture–false questions, 21
workplace-based, 60, 70, 73–74, 190–191

formula scoring (negative marking), 20, 161

Galton, Sir Francis, 142–143
gender differences, 177–178, 193
generalisability theory (GT), 154, 156

General Medical Council (GMC)
appraisal, 133

assessment practice standardisation, 187–188
international graduates, 188–189
national testing, 188
performance assessments, 70, 190–191
performance data, 197
re-validation, 190, 191
gender differences, 177–178, 193
geographical mobility, 193, 194

global rating scales, 51
GROW(N) model, 128, 129
guessing, 161

hawks see examiners, leniency/severity
healthcare changes, 186

academic, 191–192, 196–197
economic, 192–193
regulatory, 187–191

sociological, 193–194
healthcare institutions, professionalism, 91–92
healthcare teams, simulations, 64, 195

high fidelity simulations, 62–63
OSCE, 55–56
reason for, 60, 61–62
high stakes assessments see summative assessments

history taking

Osler's, 40–42
OSLERA, 40–42
Hofstee method, standard setting, 103, 104

illness, student, 170
images, use in assessments, 196

institutional phenomenon,
professionalism as, 91–92
interdisciplinary simulations, 64
international healthcare professionals, 188–189, 193
Internet, on-line assessments, 109–111, 135–136, 137, 194–195, 196
interpersonal phenomenon, professionalism as, 91
interpersonal skills questionnaire, 132
inter-rater reliability, 153–154, 162, 175, 176
item response theory (IRT), 109, 155, 156
Johari’s window, 121
key feature problems (KFPs), 28–30, 31–32
knowledge assessment, 16–36
assessable attributes hierarchy, 6–8
case-based discussion, 68
constructed answer questions, 16–17
see also essay-type questions
dissertations, 32–35
essay-type questions, 17–20, 85, 111
key feature problems, 28–30, 31–32
Miller’s pyramid, 8, 9
multiple-choice questions see multiple-choice questions
portfolios, 35–36
professionalism, 84–85
projects, 32–35
script concordance items, 30–32
selected answer questions, 16, 17
see also multiple-choice questions
Kolb’s learning cycle, 61
Kuder–Richardson 20 (KR 20), 151
language
feedback, 120, 122
students’ competence, 179
testing, 189
learning
assessment for, 4, 60, 61, 95, 119
assessment–teaching alignment, 5, 45, 105–106, 107, 110, 189
factors in student failure, 178–181
feedback, 116, 119–126, 134–135
formative–summative assessment balance, 95, 119–120
Kolb’s learning cycle, 61
portfolios for, 35
professionalism, 78–79, 80, 83, 84
simulations for, 61–66, 108, 195
workplace-based assessments, 60, 70, 73–74, 190–191
learning environments, assessment, 91
learning resources, on-line, 194
Likert scales, WPBAs, 67, 68
Log book, 35, 71–72
long case examination, 39, 40
see also objective structured long examination record
mannequins, 61, 62, 63
marking, 168–169
anonymised, 169, 176
blind double, 169
compensated marks, 170
computer-assisted, 110–111
condoned failure, 170
certainty-based, 27–28, 161
double, 169
error-reduction strategies, 176, 177
essay-type questions, 17–18, 19, 111
leniency/severity, 10, 50, 173, 174–176
long case examination record, 40
mark capping, 170
moderation, 169, 172
negative, 20, 161
OSCE, 43, 47, 51–54, 82, 110, 174, 175
OSLER, 40–42
projects and dissertations, 34–35
threats to test validity, 161, 162
see also scoring systems; standard setting
Marlow–Crowne social desirability scale, 161
Membership of the Royal College of Physicians (MRCP) 145, 149n
mentors, 182
meta-cognitive skills, 178
meta-feedback skills, 126, 134–136
Miller’s pyramid of competence, 8, 9, 59, 60, 79, 80
mini clinical evaluation exercises (mini-CEXs), 66–67, 174
mobile technology, 194
moderation of assessments, 169, 172
modified Angoff methods, standard setting, 54, 102
modified essay questions (MEQs), 17, 18–19
multiple-choice questions (MCQs), 17, 20–28
computer-based cases of the month, 135–136, 137
computer-based cheating detection, 164
computer-based item generator, 109
confidence assessment, 27–28
Nedelsky method standard setting, 101
optical marking, 110
professionalism, 84–85
scoring, 20, 27–28, 101, 161
types, 20, 21–23
writing, 24–27
multi-source feedback (MSF), 9, 130–132
professionalism, 10, 87

national testing, 188
near misses, 84
Nedelsky method, standard setting, 100–101
negative marking, 20, 161
NHS budget allocations, 192–193
norm referenced (relative) methods, standard setting, 3, 52, 99–100, 103
objective structured clinical assessment of technical skills (OSATS), 39, 44
objective structured clinical examination (OSCE), 39, 42–56
assessor severity, 175
assessors’ judgement process, 173–174
assessor training, 48–50
blueprinting, 45
candidate instructions, 45, 47
developing OSCE stations, 45–48, 82
feedback, 50
guidance for assessors, 45, 47, 48
marking, 43, 47, 51–54, 82, 110, 174, 175
planning, 44–50
professionalism, 82
reliability, 153, 154, 175
scoring, 51–52, 174
sequential, 54–55
‘serious concerns’ documentation, 50
simulated patient information, 45, 47
simulated patient scoring, 52
simulated patient training, 48–49
simulation fidelity, 55–56
standard setting, 52–54
objective structured long examination record (OSLER), 39, 40–42
optical markers, 110
Ottawa Conference 2010, professionalism, 88–92

paper-based tests see written assessments; written questions
parallel forms reliability, 150
pass score setting see standard setting
patients see service user assessment; service user images; service user simulators
patient satisfaction surveys, 86–87, 131–132
peer review, 10, 87, 130–131, 134
Pendleton’s rules, feedback, 128, 129
performance assessment, 59–74
  appraisal, 132–133
  competence and, 8–9, 38, 59
consensus statements on, 73
data transparency, 196–197
log-books, 71–72
portfolios, 72–73
reason for, 60
reliability, 153–154
simulation, 59, 60, 61–66
trends, 2–3, 190–191
workplace-based, 8–9, 38, 59–60,
  66–71, 73–74, 174, 190–191
personal development portfolios see
  portfolios
personality traits, 90
physical examination
  long case examination, 40
OSCE, 42–56
OSLER, 40–42
portfolios, 10, 35–36, 72–73, 111,
  194–195
postgraduate theses, word length, 34
power relations, feedback, 122
predictive validity, 157–158
procedural skills, OSATS, 44
professionalism, 9–10, 76–92
  assessment approaches, 79–81
  definitions, 77–78
  as individual phenomenon, 90–91
  as institutional/societal phenomenon,
    91–92
  as interpersonal phenomenon, 91
learning, 78–79
observation in vitro, 81–83
observation in vivo, 83–84
Ottawa 2010 expert group, 88–92
paper-based tests, 84–86
peer review, 87
proxy measures, 88
self-assessment, 87–88
service user assessment, 84,
  86–87
teaching, 78–79
workplace-based assessments, 10,
  82–83
Professional and Linguistics Assessment
  Board (PLAB), 189
projects, 32–35
psychometrics, 141–165
  definitions, 142–143, 144
  educational assessments as, 144
  regulatory change, 189–190
  role, 142
  test reliability, 144, 145, 147–156
  test validity, 144, 145–146, 156–164
Quality Assurance Agency (QAA),
  168–170, 172–173
quality of tests see psychometrics
questionnaires
  doctors’ interpersonal skills, 132
  patient satisfaction, 86–87, 131–132
questions, exam see written questions
reflective practice
  documenting, 35, 36, 73, 85
  feedback, 124
  professionalism, 85
  simulation and, 63
regulatory change, 187–191
relative (norm referenced) methods,
  standard setting, 3, 52, 99–100,
  103
reliability of tests, 11, 39–40, 144
blueprinting, 45
classical test theory, 147–149,
  154–155
competence theory, 153, 154
computer analysis, 109–110
definition, 145
establishing and evaluating, 147–156
generalsibility theory, 154, 156
importance, 145
internal consistency, 150–151
inter-rater reliability, 153–154, 175,
item response theory, 155, 156
item statistics, 147, 151–153
parallel forms method of assessing, 150
performance assessments, 153–154
p-value (item difficulty), 152
regulatory change, 189–190
reliability coefficients, 147, 148–149, 155
small cohorts, 189–190
split-half method of assessing, 150
standard error of measurement, 151
test re-test method of assessing, 149–150
validity and, 145
written tests, 148–153
research, into assessment, 191–192, 196–198
Research Excellence Framework (REF), 191
research projects, 32–35
resits
future rules, 197
mark capping, 170
re-validation, 190–191
right-scoring, MCQs, 20
Ronseal test, 11
see also validity of tests
Royal College of Physicians 145, 149n
sampling, 106
classical test theory, 154–155
scenarios, assessing professionalism, 85–86
scoring systems
general requirements, 168
multiple-choice questions, 20, 27–28, 101, 161
OSCE, 51–52, 174
threats to test validity, 161, 162
see also standard setting
script concordance items (SCIs), 30–32
selected answer questions, 16, 17
see also multiple-choice questions
self-assessment, 10, 87–88, 133, 134
self-awareness, 63
sequential OSCE, 54–55
‘s‘serious concerns’ documentation, 50
service user assessment, 130–132
professionalism, 84, 86–87
service user images, 196
service user simulators, 45, 47, 48–49, 52, 62
assessment centres, 195
halo effect, 178
undercover, 84
short answer questions (SAQs), 20, 85, 111
significant events, 84
simulated client observation, 84
simulation(s), 2–3, 59, 60, 61–66
assessment centres, 195
computer-based, 108, 195
OSCE, 43, 45, 47, 48–49, 52, 55–56, 82
professionalism, 81–83, 84
single best answer questions (SBAs), 17, 20, 21–22
situational judgement tests (SJTs), 85
situation-specific professionalism, 91
skill (psychomotor domain), 7, 8
simulated practice, 61–66
see also competence assessment;
performance assessment
skills-based scoring, OSCE, 51–52
smart phones, 194
socially desirable answers, 161
societal phenomenon, professionalism as, 91–92
sociological changes, 193–194
specific learning disabilities (SpLD), 178–179
split-half reliability, 150
standard error of measurement (SEm), 151
sequential OSCE, 55
standard setting, 3, 94–105
absolute methods, 3, 52, 100–102
see also below compromise methods
compromise methods, 102–105
definition, 98
European Credit Transfer and Accumulation System, 169–170
judging panels, 97
need for, 94, 96
OSCE, 52–54
p-values (item difficulty), 152
relative (norm referenced) methods, 3, 52, 99–100
see also above compromise methods
test modality–method relation, 96–97, 99
station-based scoring, OSCE, 51–52
still images, 196
stress, 180
exam, 160, 162, 163
structured (supervised) learning events (SLEs), 38, 191
student failure, 169–171, 176–182
anxiety, 180
see also stress, exam
appeals, 171, 192
attendance problems, 181
compensated marks, 170
competitiveness, 180
complaints, 171
condoned, 170
ethnicity, 177–178
gender differences, 177–178
insecurities, 180
language competence, 179
learning disabilities, 178–179
learning needs, 178
mark capping, 170
meta-cognitive skills, 178
mitigating circumstances, 170
reluctance to seek help, 179–181, 182
resits, 170, 197
stress, 180
see also stress, exam
study skills, 180
support systems, 178–179, 180–182
study skills training, 180
summative assessments, 10, 95
appraisal, 133
balance with formative, 95, 119–120
competence, 38
evaluation see psychometrics
feedback, 116, 119–120, 133
feedback on, 134, 135
OSCE standard setting, 52–54
pass score setters, 97
professionalism, 80	true–false questions, 21
workplace-based assessment, 70
supervised (structured) learning events (SLEs), 38, 191
support for students, 178–179, 180–182
surgical skills
computer-based assessment, 195
objective structured assessment, 44
surveys, patient satisfaction, 86–87, 131–132
systematic bias in tests, 158–164
teaching, professionalism, 78–79
teaching–learning–assessment alignment, 5
blueprinting, 45, 105–106, 107, 110, 189
team working simulations, 64, 195
technical developments, 194–196
test anxiety (exam stress), 160, 162, 163
test quality see psychometrics
test re-test reliability, 149–150
transparency, performance data, 196–197
true–false questions, 17, 20, 21
true scores theory, 147–148, 155
undercover observers, 84
users see service user assessment; service user images; service user simulators
validity of tests, 11, 39–40, 144, 156–164
badly written questions, 160–161, 162, 163
blueprinting, 45
cheating, 163–164
construct, 146
construct-irrelevant variance, 159–164
construct under-representation, 159
convergent, 158
definition, 145–146
divergent (discriminant), 158
exam paper errors, 160
exam stress, 160, 162
guessing, 161
predictive, 157–158
regulatory change, 189–190
reliability and, 145
small cohorts, 189–190
sources of evidence for, 156–157
threats to, 158–164
values, professionalism, 90
van der Vleuten equation, 11
verbal protocol analysis, 175–176
video recordings/images, 195, 196
workforce planning, 193
workplace-based assessments (WPBAs), 8–9, 38, 59–60, 66–71, 73–74
e-portfolios, 194–195
examiners, 70–71, 174
future of, 190–191
professionalism, 10, 82–83
writing skills, 36
written assessments
cheating, 163–164, 195–196, 197–198
computer-assisted free text marking, 111
computer-based, 195–196
reliability, 148–153
small cohorts, 189–190
validity, 157–164
see also written questions
written questions
answer cueing, 20, 22, 23–25, 161
badly written, 160–161, 162, 163
exam paper errors, 160
knowledge assessment, 16–32
professionalism assessment, 84–86
question banks, 108, 109, 162
see also written assessments