Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
1. Groningen Reflection Ability Scale	Andersen O'Neill Gormsen, Hvidberg Morcke (2014)	Netherland	361 medical students from all semesters (except 8th and 10th for pragmatic reason)		Self-evaluation questionnaire	23 items on a 5-point Likert scale (1 totally disagree, to 5 totally agree). Item scores can be summed.	Translation and cultural adaptation (Dutch-Danish-English) verified by interviews with 3 students. **Internal structure:** No factor model adequately explains the variance. No ceiling or floor effect for the overall score. Floor effect for 2 items and ceiling effect for 3 items. Internal consistency: Cronbach's alpha = 0.87 Inter-item covariance: 0.22 **Relationships with other variables** Small significant difference in score between male and female. No correlation with age, level of study and extracurricular activity.	Test-retest reliability: Bland-Altman = 3.55 Confidence Interval [0.21-6.90]	
	Aukes, Geertsma, Cohen-Schotanus, Zwierstra, Slaets (2007)	Netherland	First measurement: - 538 first-year to sixth-year medical students - 38 medical teachers - 14 experienced teachers in medical skills Second measurement : - 1029 first-year to sixth-year medical students		Self-evaluation questionnaire		**Content:** Supported by the use of literature to conceive of the items, and the covering of three major aspects of personal reflection (self-reflection, empathetic reflection, reflective communication) **Internal structure** Factor analysis revealed the presence of only one factor with three dimensions. Internal consistency: - Alpha Cronbach : 0.83 for the 1st measurement and 0.74 for the 2nd measurement - item-difficulty: between 4.23 and 3.50 - item variance: between 1.04 and 0.67		10 min to administer and complete the scale. Scores are easy to calculate
2.	Boenink, Oderwald, de Jonge, van Tilburg, Smal (2004)	Netherland	(n=10 to 15 depending on the vignettes) PGY 4 medical students (who had not begun their clerkship) following an ethics course		Vignettes : Student read reflection-evoking vignettes and were asked to write between 2 and 10 reflection documents	Overall reflection was rated on a 10-point scale 1–2 Oversimplified, intolerant opinion, only emotional reaction. 3–4 Limited ⁄ restricted, narrow-viewed, one-sided reaction, mostly just 1 perspective, no weighing up or balancing, no attention paid to context. 5 More than 1 perspective, but neither balancing nor attention paid to context. 6–7 More perspectives, general as well as personal, some balancing between perspectives. 8–9 Differentiated balancing, room for dilemmas and or doubt, explicit attention paid to the patient. 10 A subtle ⁄ balanced approach, considering all relevant perspectives, weighing up of different interests, a keen eye for dilemmas and uncertainties, paying attention to the patient's viewpoint and an evaluation of one's own position and attitude.	**Internal structure** -Low variability between vignettes (r= 0.38). Correlation was much stronger when the second vignettes were provided at the beginning of the class (r= 0.75,) then at the end (r = 0.20) Internal consistency: Pearson correlation coefficients between scores on 2 vignettes were 0.35 for case 1 and 2 and 0.41 for case 3 and 4 **Relations to other variables:** Higher score for women and students with greater clinical experience.	Inter-rater reliability 2 raters rated 4 of the vignettes A sufficient inter-rater reliability was found (Pearson r= 0.53-0.94). Only those vignettes were used in following analyzes.	Time-consuming for learners
3.	Bogo, Regehr, Katz, Logie, Mylopoulos (2011)	Canada	Social Work Master's students	Schon's	Objective structured clinical examination: five scenarios of authentic social work situations played by the students and an actor followed by a 15 min interview based on reflective-dialogue with a faculty member. Students received guidelines and prompts for each phase of the model, to encourage the development of reflection and transformational learning (Bass et al., 2017).	3 dimensions with each sub-dimension evaluated on a 5-point scale: 1. Conceptualization of practice (content, diversity, and process) 2. Self-regulation (affective, cognitive) 3. Professional Development (learning, growth)	**Internal structure** Internal consistency Overall Cronbach's alpha of 0.48. Cronbach's alpha of 0.93 for each scenario **Relations to other variables:** Supported by the higher scores (marginal effect) of experienced students		Need to train actors. Time-consuming for raters due to the 15 min interview.

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
4. Carter Assessment of Critical Thinking in Midwifery	Carter, Creedy, Sidebotham (2016)	Australia	Midwives practicing and who had mentored one student in the last six months.	Bass Model of Holistic Reflection Australian National Competency Standards for the Midwife	Observation of the learners in practice	Carter Assessment of Critical Thinking in Midwifery Only factor 2 (reflection) was considered 7 items rated according to a 6-point Likert scale (1= strongly disagree to 6= strongly agree). Factor 2 includes: -Analyses own strengths and limitations in skills, knowledge and experience -Addresses own limitations in skills, knowledge and experience Initiates professional dialogue around midwifery practice -Evaluates own practice and its effect on the woman and others -Adjusts own practice based on feedback from the woman and others -Recognizes own attitudes, biases and values and their potential impact on practice -Debriefs with a professional colleague following complex situations to improve future practice	**Content** - Items were mapped against the definition of critical thinking developed by Scheffer and Rubenfeld (2000). -11 content experts rated the relevance of the items. CVI ranged from 0.73 to 1. One item was removed and the CVI for total scale was=0.97 **Internal structure** Subscale-total correlation range : 0.88 Item sub-scale correlation range : 0.45–0.87 Internal consistency Cronbach's alpha =0.90		
	Carter, Creedy, Sidebotham(2017)	Australia	100 reflective writings from midwifery students	Bass Model of Holistic Reflection Australian National Competency Standards for the Midwife	E-portfolio containing writing on clinical, personal, or professional experience. Students used the Bass Model of Holistic Reflection comprising 6 interdependent phases: self-awareness, description, reflection, influences on knowing, evaluation and learning	Carter Assessment of Critical Thinking in Midwifery (Reflection) 15 questions divided over 3 dimensions (analyses context, reasoned inquiry, and self-evaluation) and rated on a 5-point Likert scale (1 = not at all, 5 = to a great extent)	**Content** Draft items were reviewed by two experienced midwifery researchers And mapped against the consensus definitions of the habits of the mind and skills of critical thinking in nursing **Internal structure** Factor analysis : Kaiser-Meyer-Olkin value of 0.89 The three factors (analyses context, reasoned inquiry and self-evaluation) explained 72.9% of the variance. CVI ranged from 0.9 to 1 with a total index score of 0.98. Internal consistency : corrected item-total correlation = 0.54 to 0.77.The scale demonstrated good internal consistency (Cronbach alpha =0.93) Each item had a strong correlation with the total.	Interrater reliability same rating achieved by 2 raters 72% of the time, Kappa coefficient=0.43	
5.	Costa, Driessen, Silva, Campos, Costa, Donateli Cotta (2018)	Brasil	70 portfolios from health Policy graduate students	Theoretical assumptions of Lizarraga	Development of a collective portfolio in group of 6 The students were guided through steps and met with a teacher to analyze and discuss their portfolio.	5-point Likert scale based on Cotta et al. (2011;2013): comprehensive thought, critical thought creative thought, and establishing goals.	**Relations to other variables:** Observed increase with grade over time showing evidence of improvement.	Interrater reliability: - Cronbach alpha greater than 0.7 for each cognitive skill - Agreement between pair of raters= Kappa greater than 90% for each cognitive skill	
6.	Dalal, Hakel, Sliter, Kirkendall (2012)	United States	219 university students (16% taking graduate credits)		E-portfolio unstructured and unguided	6 level rubric on 6-point Likert scale (0= Not a reflection, 5= Represents a deep reflection)	**Relations to other variables** - no correlation with the credit hours earned - no correlation with the ACT scores - moderately positive relationship with grades	Interrater reliability: - ICC for 9 experienced raters =0.946, 95% CI = 0.942 to 0.950). -ICC for 4 undergraduate students as raters = 0.90, 95% CI = 0.89 – 0.91). A minimum of 2 raters can be used for an ICC=0.946.	
7.	Devi, Abraham, Kamath (2017)	India	173 medical students		Written reflection: A teaching session about phases (Reviewing the experience, Critical analysis, and Reflective outcome) and indicators of critical reflection using Koole's model. Students were asked to write reflective summaries about their research activity.	A rubric based on the criteria of the Association for Medical Education in Europe guide based on Moon (2004).	**Content** Based on two medical education experts with experience in validation.	Inter-rater reliability (2 raters): Kappa = 0.78, 95% confidence interval (0.84, 0.72)	Reflective summary graded in 5 minutes.
8. Newcastle Reflective Analysis Tool	Dempsey, Warren-Forward, Findlay (2009)			Boud's	Written reflection	NRAT (Newcastle Reflective Analysis Tool) composed of the Deep Analytic NRAT and Broad Classification	**Content** Based on Wong's model which is based on Boud's theory		NRAT-Broad Classification Some training required. NRAT Deep Analytic: Require rater training. Require raters with minimal knowledge of reflecting in the healthcare context.
	Findlay, Dempsey,	Australia	36 radiation therapy Students (36= year 1, 32=year 2 and 29 =year 3) in		Written reflection	NRAT (Newcastle Reflective		Inter-rater reliability: - Deep Analytic NRAT= Kappa of 0.49	

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
	Warren-Forward (2010).		professional placement every year for three years.			Analysis Tool) composed of the Deep Analytic NRAT and Broad Classification NRAT that can be used together or alone. The Deep Analytic NRAT is a 6-level classification of reflection: attending to feelings, association, integration, validation, appropriation, outcome of reflection The Broad Classification NRAT has three levels of reflection (non-reflector, reflector, critical reflector).		- Broad Classification NRAT =Kappa 0.67 Deep Analytic NRAT (4 raters) : - level 1: moderate absolute agreement (69.1%), and very high consensus agreement (89.7%). -level 2 and 3: poor absolute agreement (26.8% and 50.5%) and good consensus agreement (75.3% and 85.6%) - level 4-5-6: absolute (85.6%–90.8%) and consensus agreement (93.8%–99.0%) were very high. Broad Classification NRAT: -three levels showed good absolute (75.3%–81.4%) and excellent consensus (89.7%–92.8%) agreement	
	Findlay, Dempsey, Warren-Forward (2011)	Australia	5 intern radiation therapists		Written reflection: Questions guiding reflective writing about post-workshop reflection, significant clinical event reflection, post-journal reading reflection.	NRAT (Newcastle Reflective Analysis Tool)		<u>Inter-rater reliability (two raters)</u> - Deep Analytic NRAT = 75.0 to 83.3%, Kappa value = 0.47 to 0.59 - Broad Classification NRAT =97.3 to 100.0%) and Kappa =0.94 to 1.00.	
9. The Self-Reflection and Insight Scale	Grant, Franklin, Langford (2002)		288 and 121 undergraduate psychology students for the factor-analysis and for convergent validity		Self-administered questionnaire	The Self-Reflection and Insight Scale Six-point scale (1= strongly disagree, 6= strongly agree) developed initially with 3 categories and 30 items	**Internal structure:** Factor analysis: - Two-factor scales of 20 items (12 items on the factor Self-Reflection and 8 on the Insight scale) - Two-factors explained 56% of the variance Internal consistency: - Self-reflection scale: Cronbach alpha= 0.91 - Insight scale: Cronbach scale =0.87 - Nonsignificant correlation (r=-0.03) between the two scales **Relations to other variables** - Positive correlation between the private self-consciousness scale, anxiety and stress and the Self-reflection scale - No correlation between cognitive flexibility, self-regulation, depression or alexithymia and the Self-Reflection scale. -Negative correlation between the private self-consciousness scale and the Insight scale - Negative correlation between depression, anxiety and stress and alexithymia and the Insight scale. - Positive correlation between cognitive flexibility, self-regulation and the Insight scale.	<u>Test-retest reliability</u>: over 7 weeks - Self-reflection scale = 0.77 - Insight scale = 0.78	

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
The Self-Reflective Writing Scale	Ho-wai So, Benett-Levy, Perry, Wood & Wong (2018)	Hong Kong	36 students of a clinical psychology university program		Scale to evaluate self-reflective writing	Scale divided in two domains (personal-self domain and therapist-self domain) with four components each: Personal-self domain: - Self-awareness - Attitudes - Interpersonal skills - Personal knowledge and experience Therapist-self: - interpersonal declarative knowledge and/or skills, - conceptual knowledge and/or skills, - technical knowledge and/or skills, - plans procedures and skills Self-reflective writings were noted according to: clarity, depth, comprehension, specificity and precision on a scale from 1 (Basic) to 3 (Advanced).	**Internal structure** Factor analysis Significant within-person difference between the two domains. **Relations to other variables** Convergent validity Level of self-reflectivity correlates with interpersonal skills in clinical placement. Reverse correlation between personal-self reflectivity and relationship building	Interrater reliability ICC for personal self = 0.6 ICC for therapist self = 0. 61	
10.	Jensen, Joy (2005)	United States	20 junior baccalaureate nursing students (60 journals)	Meizrow	Written reflecton: Students received a lecture about the seven levels of reflection of Meizrow. At three times during their class, they were asked to write about an uncomfortable situation that had changed their way of thinking.	Score obtained by multiplying the number of items score by the level scored. Seven level of reflection of Meizrow (1981): Level 1: Reflectivity. Level 2: Affective Reflectivity. Level 3: Discriminant Reflectivity. Level 4: Judgmental Reflectivity. Level 5: Conceptual Reflectivity. Level 6: Psychic Reflectivity Level 7: Theoretical Reflectivity.		Interrater reliability: - Number of items scored: Pearson r = 0.74 - Total score: Pearson r = 0.63	
11. Reflection Questionnaire	Kember, Leung, Jones, Loke, McKay, Sinclair, Tse, Webb, Yuet Wong, Wong, (2000)	Hong Kong	303 students from 8 classes of a health science faculty	Meizrow	Self-rated questionnaire	4 scales each measured by 4 items: - Habitual action - Understanding - Reflection - Critical reflection	**Content** Developed based on reflective thinking literature **Internal structure:** Factor analysis: Good fit with a 4-factor model Internal consistency: Cronbach alpha: - Habitual action = 0.621 - Understanding= 0.757 **Relations to other variables** - Postgraduates scored significantly higher on reflection and critical reflection than undergraduates. - Reflection = 0.631 - Critical reflection = 0.675		
	Lethbridge, Andrusyszyn, Iwasiw, Laschinger, Fernando (2013)	Canada	303 of students (including nursing students) from eight classes (six undergraduate, one post diploma and one Master's) in a health sciences faculty				**Internal structure** Factor analysis: The CFA results verified a 4-factor model (w2 = 179.3, df = 100, CFI = 0.903), Internal consistency: Cronbach's alpha of 0.62, 0.76, 0.63, and 0.68 for the four scales, respectively		
			538 baccalaureate nursing students				**Internal structure** Factor analysis : model fit supporting the hypothesis that the four first-order factors can be combined in a second-order structure Internal consistency: Cronbach alpha: 0.58 to 0.85		

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
	Leung, Kember (2003)	Hong-King	402 undergraduates from a health science faculty				**Relations to other variables** Convergent validity: Correlation with deep learning according to the Study Process Questionnaire and the Study Process Questionnaire Revised Divergent validity: No significant correlation with Habitual action scale No correlation with surface learning according to the Study Process Questionnaire and the Study Process Questionnaire Revised		
	Kember, Jones, Loke, McKay, Sinclair, Tse, Webb, Wong, Wong & Yeug	Hong Kong	Student of nursing, occupational therapy, physiotherapy and radiotherapy	Mezirow	Journal writing			Inter-rater fidelity ICC for 3 reflective writing = 0.64 ICC for 9 reflective writing evaluated by 4 raters = 0.74	
	Wallman, Lindbland, Hall, Lundmark & Ring (2008)	Sweden	Pharmacy interns	Mezirow	Written reflection	6 levels of reflection: (1) habitual action (2) thoughtful action (3) introspection (4) content reflection (5) process reflection (6) premise reflection	**Sensitivity:** Evidence that changes can be detected.	Inter-rater fidelity Cohen's kappa for 182 written evaluations assessed by 2 raters = 0.63	Mean time to rate a copy = 3 minutes
	Tutticci, Coyer, Lewis & Ryan (2017)	Australia	346 final year nursing students	Mezirow	Questionnaire following a clinical situation simulation	4 scales each measured with 4 items: - Habitual action - Understanding - Reflection - Critical reflection	**Internal structure** Factor analysis Factor analysis showed lack of construct validity. Internal consistency Fair to good (0.711) **Relations to other variables** Convergent validity: No evidence of convergent validity Discriminatory validity: No evidence of discriminatory validity		

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
13.	Learman, Autry, O'Sullivan (2008)	United States	32 residents in an obstetrics and gynecology program		Written reflection	6-point rubric 1. describes procedure/case/setting without mention of lessons learned; 2. states opinions about lessons learned unsupported by examples; 3. superficial justification of lessons learned; 4. reasoned discussion well-supported with examples regarding challenges, techniques, and lessons learned; 5. analyzes factors from experience that contribute to progress; and 6. includes justification for the strategies that were used and evidence for effectiveness	**Internal structure** Internal consistency Low (Cronbach's alpha = 0 .40) for all 6 exercises combined adequate (Cronbach's alpha = 0.62) for 5/6 exercises Construct validity Senior learners scored higher **Relations to other variables** Convergent validity Correlation between Reflection scores and professionalism and communication (r= 0.36 to 0.37) No correlation with medical knowledge	Inter-rater reliability r= 0.89 for 18 copies corrected by 2 raters	
14.	Kennison, Misselwitz (2002)	United States	17 sophomore and junior nursing students		Journaling: describing situation in which they used their critical skills.	8-point Likert scale for each of the 6 dimensions of the Faccione's theory (interpretation, analysis, inference, evaluation, explanation and self-regulation) Baker's four-step model was included		Inter-rater reliability No consensus between the raters was reach.	No training of the raters
15. Student Assessment of Reflection Scoring rubric	Koole, Dornan, Aper, De Wever, Scherpbier, Valcke, Cohen-Schotanus, Derese, (2012)	Belgium	273 medical students		Video-case: 4 interactive video-cases with 6 interruptions when students had limited time to answer a question. After the video, students had to answer 6 questions to reflect on their experience.	Student Assessment of Reflection Scoring rubric measured on a 4-point scale - 2 questions about awareness - 2 questions about understanding - 2 questions about future action	**Internal structure** Validity supported by the large variation in scores	Interrater reliability: Kappa coefficient = 0.88 (2 raters)	For learners: about one hour to view the case and reflect upon it. For raters : 30 min training 3 hours to score 40 students. To develop the video: 5 hours training for experienced simulated patients +time to create 15-20 min video-cases.
16.	Lia & Wang, 2016	Taiwan	86 medical students		Reflective writing	6 domains: -focus & context structure, -ideas -voice & point of view - critical thinking & representation - depth of reflection on personal growth -language & conventions.		Interrater reliability: Chronbach's alpha between 0.82 and 0.88	
17.	Lewis, Virden, Hutchings, Bhargava (2011)	United States	Applicants for psychology doctoral program		Verbal reflection: Qualifying examination to determine the advancement of doctoral candidates. Students were asked to reflect verbally with two faculty members on their development and performance on 35 task accomplished during their graduate program. They received a lecture about the qualifying examination.	Scoring Rubric ranging from 1 to 6 1 : (Beginning/Novice level) 3: Developmentally expected Level 6 : Sophisticated/Advanced Level	**Content** Experts responding to a survey indicated that: - all believed that students passing the QE demonstrate self-reflective practices at a level appropriate for advancement to doctoral candidacy - 80% indicated that a student passing the QE understands the competence areas required in professional psychology		90 minutes assessment with a minimum of two faculty members for each learner
18.	Lucas, Bosnic-Anticevich, Schneider, Bartimote-Aufflick, McEntee Smith (2017)	Australia	43 pharmacy students	Boud	Written reflection answering to the following questions: 1. What were your thoughts, feelings and beliefs towards the task, and were they altered by performing the task? 2. How did you prepare for this task? 3. What happened during the task? 4. Your strengths or skills that you found applicable to the task? 5. Did you encounter any problems (personal, weaknesses, barriers addressed) whilst undertaking the task and if so, how did you overcome them? 6. What have you learned as a result of this task and how may it benefit you in the future?	Each reflection was grade based on 7 stages: (1) returning to the experience, (2) attending to feelings, (3) association, (4) integration, (5) validation, (6) appropriation (7) outcomes of reflection. Each stage was graded "0" (non-reflector), "0.5" (reflector), or "1" (critical reflector) and scores are added.		Inter-rater reliability: -for the overall reflective statement ICC = 0.81 (95% CI 0.61–0.90) - for average score for each stage ICC = 0.89 (95% CI 0.83–0.93),	

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
19.	McEvoy, Pollack, Dyche & Burton, 2016	USA	186 second year medical student		3 written questions 1. What progress do you feel you have made with the personal goal you set at the beginning of the year? How do you measure this progress? Is there more that you might do to advance the goal? 2. How did you deal with the personal challenges you anticipated in pursuing this goal? Were there pivotal events or people that helped you? 3. What do you see as an appropriate goal for next year and why?	Level 1: No reflection Level 2: Goal setting without reflection Level 3: Reflection that falls short of accessing feelings Level 4: Cognitive/affective reflection			
20.Reflection Evaluation for Learners' Enhanced Competencies Tool And 21. Holistic Reflection-on-Action rubric	Miller-Kuhlmann, O'Sullivan, Aronson (2016)	United States	56 third-year medical students with little prior exposure to reflection	Narrative reflective framework	Written reflection: Students were asked to choose an experience during their clerkship that had an impact on them or on the medical profession and to write about it.	Comparison of 2 evaluations: - Reflection Evaluation for Learners' Enhanced Competencies Tool (REFLECT), based on the narrative reflection Framework. Analytic rubric of 5 dimensions of reflection which are rated on a 4-level scale (habitual action to critical reflection) - holistic Reflection-on-Action rubric comprising one dimension (reflective ability) rated on 6 hierarchical levels.	Reflection-On-Action: **Internal structure** Scores had a bottleneck effect at level 3. Factor analysis: Some students demonstrated higher level skills without demonstrating basic levels, which is inconsistent with a one-dimension assessment. **Relations to other variables** Increased score for both evaluations when an educational intervention was offered REFLECT: **Internal structure** Vulnerable to a ceiling effect. **Relations to other variables** Increased score for both evaluations when an educational intervention was offered	Inter-rater reliability Raters were able to achieve higher reliability in less time for Reflection-on-Action.	Reflection-on-Action: initial training 2 hours. REFLECT: initial training 4 hours. Two more hours were needed due to discrepancy between raters about specific domains.

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
22. Level of Reflection-on-Action Assessment	Padden (2013)		Nursing students	Boud	Journaling: Students wrote a reflective journal about their anticipation of their clinical experience. A series of questions were available to guide their journal based on Nielsen's, Stragnell's, and Jester's (2007) Guide for Reflection.	Level of Reflection-on-Action Assessment (LORAA) is based on the six levels of reflection-on-action by Boud et al. (1985): Level 1. Student describes the incident, including what was initially noticed. Level 2. Student describes beliefs and feelings about the incident. Level 3. Student relates the incident to similar past encounters, experience, or theoretical knowledge and includes additional data collected and any assistance received with problem solving. Level 4. Student states analysis(es) and interpretation(s) of the situation. Level 5. Student provides information regarding responses to the situation, actions, or interventions taken, or not taken, and why. A statement or statements regarding stressors experienced is included, along with recognition of new perceptions of previously held beliefs and knowledge. Student provides information reporting the response of the patient or other individuals involved in the situation and any subsequent actions taken. Level 6. Student provides a statement as to what was learned from the situation. Student provides information identifying the need for additional knowledge and skills in handling similar situations in the future, as well as changes in values or feelings as a result of the experience. A level was attributed to each journal.	**Content** Validated by three experts in reflective learning	Interrater reliability : - examined journals by 5 graduates of a nursing program. For 4/5 entries = 67 to 100% -inter-rater reliability = 0.80	
23.	Pee, Woodman, Fry, Davenport (2002)	United Kingdom	26 dental therapy students	Boud	Worksheet: Students were asked to answer the following: - Briefly describe what happened - Describe your feelings at the time this happened - Why do you consider this experience to be worthy of reflection? - What strengths in your clinical practice did this experience demonstrate? -What learning needs did this experience reveal to you? -Which one learning need, disclosed by this experience, do you wish to address as a priority? - Decide exactly what you would like to achieve in relation to your selected learning need, before completing Target Setting overleaf - Target setting (target, action, resources, success criteria) -Review: what progress have you made in reaching your target, what impact had this learning had upon your professional practice, is there any further learning that you wish to undertake in this area?	Comparison of Hatton and Smith level, John's questions and the opinion of colleagues about the presence of reflection Hatton and Smith: 1 'descriptive' 2 'descriptive reflection' 3 'dialogic reflection' 4 'critical reflection'	**Process response:** Assessment of students' reflections made on the basis of their writing may not be accurate. There is a risk that reflection takes place but cannot be detected in the text. **Relations to other variables** Both students (1 and 5) whose worksheets were examined were deemed to have reflected by all the peer raters. Student 1 was judged to have reflected 'to some extent' by 44% of peer raters, and to have reflected 'thoroughly' by 56% of peer raters.	Inter-rater reliability: Using Johns' questions The researchers agreed that a question either was or was not addressed in 182 cases and disagreed in 70 cases, giving IJA of 72% for this method Using Hatton and Smiths' criteria The researchers agreed that a level either was or was not exhibited in 48 cases and disagreed in 8 cases, giving IJA of 86% for this method	

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
24.	Pitts, Coles, Thomas, Smith (2002)	United Kingdom	12 prospective general practice trainers		Portfolio	7 criteria were coded from 0–5. Scores of 0–2 were considered as 'refer' and 3–5 as 'pass' 1. Reflective learning process: Demonstration of learning, proof of use, change 2. Awareness of 'present state', willingness to learn. Learning experiences, hopes and expectations, identification and definition of educational needs shown 3. Recognition of effective teaching behaviors. Listening, questioning, identifying 'wants' and 'needs', defining and agreeing agenda, reflective teaching, summarizing, evaluating 4. Identifying with the learner. Recognizing uncertainty, acknowledging ignorance, learning together 5. Awareness of educational resources Literature, peers, mentor, courses, own learning 6. Drawing conclusions on the future. Overall gain, career development etc. 7. What evidence of use of the guideline/framework was there?		Inter-rater reliability (8 raters): On 12 portfolios: - 7 succeeded unanimously or almost unanimously - 9 succeeded according to more than half raters - 2 were referred by more than half raters Agreement range between 'slight' to 'fair' kappa= 0.26. Using four pairs of raters instead of eight increased the agreement coefficient to kappa= 0.50.	
25.	Plack, Driscoll, Blissett, McKenna and Plack (2005)	United States	27 physical therapy students (43 journals)	Boud Meizrow Schon	Journaling: Students had a 1.5-hour session of discussion about reflective practice and its importance. They were provided with definitions and questions to use if they had difficulty starting their journal.	Journals were examined for evidence of the 9 elements of reflection (reflection in action, reflection on action, reflection for action, content reflection, process reflection, premise reflection, returns to experience attends to feelings, and evaluation of the experience). Overall, journals were rated as showing no evidence of reflection, evidence of reflection, or evidence of critical reflection.		Inter-rater reliability (3 raters): - agreement in 65.1% to 93.0% of journals - Intra-class correlation coefficient: 0.03 to 0.72 (0.41 to 0.72 when «returns to experience» and «attends to feeling» were removed).	Minimal training
26.	Plack, Driscoll, Marquez, Cuppernull, Maring, Greenberg (2007)	United States	21 third-year medical students on a 6-week pediatric clerkship (308 journal entries)	Bloom taxonomy	Journaling on their clerkship	Based on Bloom's modified Taxonomy: 1) knowledge and comprehension 2) application and analysis 3) synthesis and evaluation Global entries are graded		Inter-rater reliability: - Reliability between pairs of raters for each journal entry: 78.2%–100%, kappa =0.57 0.04 to 0.73 _ 0.04 - Reliability between rater pairs for the highest level in each journal (kappa statistic = 0.52) 0.04 between raters 1 and 2; 0.52 _ 0.04 between raters 2 and 3; and 0.58 _ 0.04 between raters 1 and 3. - interrater reliability among all 3 raters for the highest level in each journal, ICC= 0.79 - interrater reliability for level 1: 1.0 (95% CI, 1.00 −1.00) - interrater reliability for level 2: 0.67 (95% CI, 0.61− 0.72) - interrater reliability for level 3: 0.62 (95% CI, 0.56–0.68)	

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
27.	Rees, Sheard (2004)	United Kingdom	100 second year medical students		Reflective portfolio: 800-word reflection completed with documentary evidence collected during the Communication skill class.	Participants are rated on 5 areas. One area was critical reflection based on the following criteria: 3 : Outstanding evidence of critical reflection Students reflect critically on their communication skills and spend time within the portfolio discussing why they experienced the problems they did. 2: Some evidence of critical reflection Students reflect critically on their communication skills but spend little time within the portfolio discussing why they experienced the problems that they had. 1: Evidence of reflection, but is not critical. Students describe their communication skills but only discuss the aspects of their communication skills that were good. Students do not reflect critically on their communication skills. 0: No evidence of reflection Students do not reflect on their communication skills (strengths or weaknesses). Here, the students discuss communication skills theory only and make no reference to their own communication skills.		Inter-rater reliability: Moderate: 95% CI = [0.239, 0.584]	
28.	Rees, Shepherd, Chamberlain (2005)	United Kingdom	121 first-year medical students		Written portfolio: Student must reflect critically on their performance and create a learning plan considering three areas: academic, personal and interpersonal.	Raters gave a global grade to the reflection on a 3-point scale: unsatisfactory, satisfactory or excellent that was converted into a numeric scale from 1 to 3. Raters then choose a statement corresponding to student reflection in a grid including 6 areas: written communication skills; level of critical reflection; use of documentary evidence; use of relevant theory/literature; identifying learning goals; and achieving	**Relations to other variables** - Significant negative relationship (r = –.512) between the critical reflection assessment and the average grade of the portfolio content. - Significant relationship between the critical reflection assessment and the written assessment (scientific report) for the 1st analysis (r = .273) but not the second (r = .16). Variance : - 69.2% was explained by the interaction student-case - 25% was explained by students' ability - 5.8% was explained by occasion	**Test-retest** 69% of the variance in grades over the two occasions was explained by the interaction between students and case 5.8% of the variance was explained by the occasion	
29.	Root, Waterfield (2015)	United Kingdom	22 first year Master's in pharmacy program		Written reflection : 650-750 words about an interprofessional activity including exercises and case discussion with students of other professions.	3 outcomes : 1.Analysis and Evaluation 2.Identification of personal learning 3.Action plan for personal development was graded as Beginning', 'Developing', 'Competent' or 'Proficient'	**Content** Lack of clarity in the assessment obscured assessment of the reflexive skills of the students. **Process response** The format of the reflection inhibited the linking of activities necessary for a reflective cycle.		
30. Reflection in learning scale	Sobral (2000)	Brasilia			Self-reported questionnaire	10 item self-rated questionnaire to evaluate self-reflection in learning using 7-point scale from 1 «never» to 7 «always»	**Internal structure** Factor analysis: 2 dimensions: integration and monitoring of learning Internal consistency: -alpha = 0.81 **Relationships with other variables** Significant correlation with Mitchell's analogous reflection scale and with measures of self-monitoring and learning.	Test-retest = 0.63 after 3 months	
	Sobral (2001)	Brasilia	196 third semester medical students			Reflection in learning scale 14 items	**Internal structure:** 48% of the variance in the Reflection in learning scale was explained by the results in Course Valuing, Meaning Orientation and perceived personal efficacy in self-reflection. **Relationships with other variables**		

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
							Significant and positive correlation between score in the Reflection in learning scale at the end of the term and third term GPA.		
	Sobral (2005)	Brasilia	147 third year medical students			Reflection-in-Learning Scale 14 items	**Internal structure:** <u>Variation:</u> At the end-of-term: 56% was explained by personal efficacy, start of term Reflection in Learning Scale score, third term Course Valuing Index score. Gender, first-year Course Valuing Index score and first-year GPA did not explained significantly variance. <u>Factor analysis:</u> Dominant first-factor start-of-term (variance: 34.3% first factor vs. 53.2% total) end-of-term (variance: 40.8% first factor vs. 57.4% total) <u>Internal consistency:</u> Reliability analysis showed good internal consistency numbers for both start-of-term (alpha=0.84) and end-of-term (alpha=0.88) measures. item-total correlation coefficients A the beginning of the term : stronger for item 5 (r=0.631) and item 1 (r=0.613). At end-of-term: stronger for item 12 (r=0.670), item 5 (r=0.646), item 11 (r=0.637), item 7 (r=0.620), and item 4 (r=0.606). The weakest was for item 13: at start (r=0.353) and end-of-term (r=0.420). Inter-item correlation: positive for all item pairs, (start-of-term (mean=0.30); end-of-term (mean=0.36)) Paired-case start–end comparisons: Significant (p<0.001) correlation coefficients for all cases. <u>Sensitivity/specificity</u> With a 76 point cut-point: - sensitivity = 69% - specificity = 87% **Relationships with other variables** Total cores were associated with scores in the «knowledge structure in memory» and «flexible thinking» dimensions of the Diagnostic Thinking assessment. Scores at the end-of-semester correlate with change in GPA. Gender or age didn't influence the score	<u>Test–retest reliability</u> 6 months interval; r=0.709 18 months interval: r=0.515 <u>Standard error of measurement:</u> 5.4 to 5.8%	
31. Modified Mayo Evaluation of Reflection on Improvement Tool	Tully, Murphy, Fioratou, Chaudhuri, Shaw, Davey (2018)	Scotland	322 medical students	Transformative learning theory	Group written reflection on an incident that happened to someone else. A discussion occurred with a Senior Nurse or Consultant. Students were then asked individually to write a reflection on the incident based on the structure of mMERIT.	Modified Mayo Evaluation of Reflection on Improvement Tool (mMERIT) 3 factors graded from 0 to 7 Factor 1: Personal Characteristics - Quality of reflection on doctors practice - Sufficient details to delineate contributing factors - Relevant new behaviors were proposed - Doctor questioned their readiness to practice. - Multiple options for personal change were considered. - Contributing personal factors were identified - Next steps towards personal change were identified. Factor 2 : System characteristics	**Relations to other variables** mMERIT showed high internal consistency Cronbach's alpha = 0.95) There was no gradient of grade depending on the time of the year.	<u>Inter-rater reliability:</u> Overall reliability for the three raters G =0.87 The reliability for each question between raters varied from G = 0.71–0.87 Inter-rater reliability results of (G = 0.56–0.79) was achieved between one and three raters	

Tool name	Author (Date)	Country	Population	Theoretical framework	Activity and nature of assessment tool	Criteria of evaluation	Validity	Fidelity	Feasibility/acceptability
						- Quality of reflection on the institution or wider health care system. - Current institutional practice or system was questioned - Contributing system factors were identified. - Relevant changes to the system were proposed - Next steps towards system change were identified - Multiple options for system change were considered Factor 3 : Problem of merit - Event was patient-centered - Potential for event to affect other patients - Event could cause negative clinical impact - Overall problem of merit - Quality gap established from standards and guidelines (local or national)			
	Wittich, Beckman, Drefahl, Mandrekar, Reed, Krajicek, Thomas (2010)	Australia	50 medical residents		Written reflection	Mayo Evaluation of Reflection on Improvement Tool who had four domains and 19 items rated according to a 4-point scale (no, somewhat, almost, and yes or bottom, second, third, top quartile).	**Content** content developed by a panel of experts **Internal structure** Factor analysis: -reveal 3 factors (personal characteristic of quality improvement, reflection on system characteristic and Problem of Merit). -Factors accounted for 86% of the variance Internal consistency: -Overall Cronbach's alpha =0.93 - Factor 1: Cronbach's alpha =0.91 - Factor 2: Cronbach's alpha =0.91 - Factor 3: Cronbach's alpha =0.83	Inter-rater reliability: - intraclass correlation coefficient: 0.73-0.89	
32. Reflective Ability Clinical Assessment	Tsingos-Lucas, Bosnic-Anticevich, Smith (2016)	Australia	199 pharmacy students		Video-case: Clinical scenario, a video podcast and a reflexive statement. Students were asked to reflect on how to overcome challenges, how they may have changed or developed and how they could improve.	Reflective Ability Clinical Assessment 16 items rated on a 5-point Likert scale			Good acceptability
33.	Williams, Sundelin, Foster-Seargeant, Norman (2000)	Canada	58 physiotherapy students during their third trimester. Students all had a baccalaureate in another domain before entering this program.	Boud	Journaling: 2 entries per week in a reflection journal about their adaptation to problem-based and self-directed learning.	The criteria for grading the journals were: 0. Non-reflection 1. Describe the learning event, issue or situation. Describe prior knowledge, feelings or attitudes. 2. Analyzes/re-evaluates the learning event, issue, or situation in relation to prior knowledge, feelings or attitudes. 3. Verifies/confirms the learning event, issue or situation in relation to prior knowledge, feelings, or attitudes. 4. Relates 1,2 and 3 above to gain a new understanding of the learning event, issue or situation. 5. Indicates how the new learning event, issue or situation will affect future behavior. Determines the clarification of an issue, the development of a skill, or the resolution of a problem.	**Internal structure** Internal consistency Split-half reliability: moderate correlation (r= 0.74) Variance: - significant interaction time-rater but the mean difference was small (less than 1/6 point)	Inter-rater reliability (3 raters): -0.68 (95% confidence interval = 0.49-0.87) No significant differences between the first and second journals showing no evidence of learning.	
34.	Wong, Kember, Chung, Yan (1995)	Hong Kong	45 registered nurses in continuing education.	Boud Meizrow	Written reflection: Students received a two-hour lecture on reflection. They created and implemented a teaching plan and had to write a reflection on their experience.	Qualitative coding based on the following categories: 1. Attending to feeling 2. Association 3. Integration 4. Validation 5. Appropriation 6. Outcomes of the reflection	**Relations to other variables** Same level of reflection was noted when students were interviewed and offered the chance to elaborate on their reflection.	Inter-rater reliability - for attribution of a global level to the student (5 raters) coefficient=0.88 - for attribution of a level to each segment (3 raters): coefficient ranged between 0.5 and 0.75.	

Supplement Table 1: Characteristics of the assessment tools.