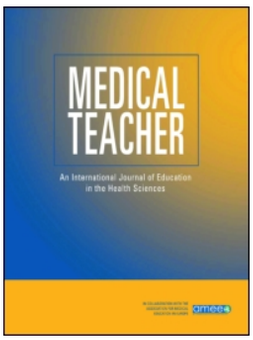

Medical Teacher

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/imte20

Psychometric validation of the reconstructed version of the assessment of reasoning tool

Satid Thammasitboon , Moushumi Sur , Joseph J. Rencic , Gurpreet Dhaliwal , Shelley Kumar , Suresh Sundaram & Parthasarathy Krishnamurthy

To cite this article: Satid Thammasitboon , Moushumi Sur , Joseph J. Rencic , Gurpreet Dhaliwal , Shelley Kumar , Suresh Sundaram & Parthasarathy Krishnamurthy (2020): Psychometric validation of the reconstructed version of the assessment of reasoning tool, Medical Teacher, DOI: 10.1080/0142159X.2020.1830960

To link to this article: https://doi.org/10.1080/0142159X.2020.1830960

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2020.1830960

Psychometric validation of the reconstructed version of the assessment of reasoning tool

Satid Thammasitboon[a,b], Moushumi Sur[a], Joseph J. Rencic[c], Gurpreet Dhaliwal[d,e], Shelley Kumar[b], Suresh Sundaram[f] and Parthasarathy Krishnamurthy[a,g,h]

[a]Department of Pediatrics, Section of Critical Care Medicine, Baylor College of Medicine, Houston, TX, USA; [b]Department of Pediatrics, Center for Research, Innovation and Scholarship in Medical Education, Texas Children's Hospital, Houston, TX, USA; [c]Department of Medicine, Boston University School of Medicine, Boston, MA, USA; [d]Department of Medicine, University of California San Francisco, San Francisco, CA, USA; [e]Medical Service Department, San Francisco VA Medical Center, San Francisco, CA, USA; [f]Department of Administration, Alfred Lerner College of Business & Economics, University of Delaware, Newark, DE, USA; [g]Department of Marketing and Entrepreneurship, C.T. Bauer College of Business, University of Houston, Houston, TX, USA; [h]Department of Anesthesiology and Pain Medicine, University of Texas Medical Branch, Houston, TX, USA

ABSTRACT

Background: Assessing learners' competence in diagnostic reasoning is challenging and unstandardized in medical education. We developed a theory-informed, behaviorally anchored rubric, the Assessment of Reasoning Tool (ART), with content and response process validity. This study gathered evidence to support the internal structure and the interpretation of measurements derived from this tool.

Methods: We derived a reconstructed version of ART (ART-R) as a 15-item, 5-point Likert scale using the ART domains and descriptors. A psychometric evaluation was performed. We created 18 video variations of learner oral presentations, portraying different performance levels of the ART-R.

Results: 152 faculty viewed two videos and rated the learner globally and then using the ART-R. The confirmatory factor analysis showed a favorable comparative fit index = 0.99, root mean square error of approximation = 0.097, and standardized root mean square residual = 0.026. The five domains, hypothesis-directed information gathering, problem representation, prioritized differential diagnosis, diagnostic evaluation, and awareness of cognitive tendencies/emotional factors, had high internal consistency. The total score for each domain had a positive association with the global assessment of diagnostic reasoning.

Conclusions: Our findings provide validity evidence for the ART-R as an assessment tool with five theoretical domains, internal consistency, and association with global assessment.

KEYWORDS
Clinical reasoning; diagnostic errors; competency-based assessment; clinical preceptor; feedback

Introduction

In recent years, the medical community has highlighted the prevalence of diagnostic errors in medicine. In 2015 the National Academy of Medicine called for improved teaching methods related to the diagnostic process and clinical reasoning (Balogh et al. 2015). To improve their reasoning skills, learners need to receive accurate context-specific feedback on their performance. However, validated tools to assess diagnostic reasoning in the clinical arena have been lacking. To address this need, the education committee of the Society to Improve Diagnosis in Medicine (SIDM) developed a theory-informed, 5-domain Assessment of Reasoning Tool (ART) (Thammasitboon et al. 2018) that provides an explicit structure for assessment of diagnostic reasoning using contemporary clinical reasoning terminology to facilitate formative feedback between teachers and learners (Supplemental Content Figure 1).

Given that the construct of clinical reasoning is broad and multi-dimensional (Young et al. 2018), assessment methods can vary significantly in the tasks of clinical reasoning that they evaluate (Gruppen 2017; Daniel et al. 2019).

Practice points

- Multiple observation tools for clinical skills have been published but none are specific to the diagnostic reasoning process.
- The Assessment of Reasoning Tool is a theory-informed, behaviourally anchored rubric that allows clinical teachers to assess diagnostic reasoning and structure feedback conversations in five domains.
- A deconstructed version of the ART (ART-R), is a 15-item, Likert Scale assessment tool derived from the behavioural descriptors of the original ART rubric.
- Psychometric evaluation provides validity evidence pertaining to internal structure and relationship to other variables of the ART-R.
- The ART and the ART-R are complementary; both can be used for assessing leaners' competence in diagnostic reasoning.

CONTACT Satid Thammasitboon satidt@bcm.edu Department of Pediatrics, Center for Research, Innovation and Scholarship in Medical Education and the Texas Children's Hospital, 6651 Main St. Houston, TX 77030, USA

Supplemental data for this article can be accessed here.

These methods include non-workplace-based, simulation-based, and workplace-based assessments (Daniel et al. 2019). There are existing assessment tools in literature (Kogan et al. 2009), such as the mini-Clinical Evaluation Exercise (mini–CEX) (Norcini et al. 2003; Donato et al. 2008), the Integrated Direct Observation Clinical Encounter Examination (IDOCEE) tool (Abouna 1999), Problem Representation, Background Evidence, Analysis, Recommendation (PBEAR) tool (Carter et al. 2018), and the Interpretive summary, Differential diagnosis, Explanation of reasoning, Alternatives (IDEA) tool (Baker et al. 2015). These tools either feature a global assessment of clinical reasoning or focus on broad domains of clinical reasoning, without deconstructing the steps of the process (van der Vleuten et al. 2008). Most of these tools also lack anchors and detailed descriptors of the sub-tasks of the clinical reasoning process (Norcini et al. 2003; van der Vleuten et al. 2008; Kogan et al. 2009; Baker et al. 2015; Carter et al. 2018). The ART complements these tools by providing detailed, behaviorally-anchored scales for multiple domains of diagnostic reasoning. The behavioral anchors can be used for assessment as well as formative feedback. When used across multiple clinical presentations in diverse clinical settings, ART-guided feedback can aid learners in the development of competency in diagnostic reasoning ability.

A committee of medical educators within the SIDM developed the ART by synthesizing relevant diagnostic reasoning frameworks (e.g. script theory, structural semantics) (Bordage 2007; Custers 2015), contemporary practice goals (e.g. high-value care), and error reduction strategies (e.g. metacognition). In a previous study, we established the ART's validity by using Messick's unified framework (Messick 1989), which focuses on content validity, response process, internal structure, relationship to other variables, and consequences. Content validity was established through theory-informed development as well as the integration of expert opinion, consensus, and feedback from clinician-educators. The pilot deployment of the ART with a convenience sample of clinician-educators that used the ART to assess and provide feedback to their learners during a clinical presentation provided validity evidence on the response process (Messick 1989). Clinical teachers found the tool to be straight forward and useful in providing explicit structure and shared language to guide teaching and feedback (Thammasitboon et al. 2018).

We describe here the continued accrual of validity evidence regarding the ART's theoretical framework (i.e. five domains of diagnostic reasoning) and its behavioral descriptors. For psychometric purposes, we reconstructed the ART into its 15 items; this version of the tool is termed the ART-R. This paper describes the psychometric analysis of the ART-R.

Methods

The instrument

The ART is an instrument designed to provide a structure for assessment and feedback to learners pertaining to five domains of the diagnostic reasoning process including hypothesis-directed information gathering (HD), problem representation (PR), prioritization of the differential diagnosis (PD), diagnostic evaluation in a manner that reflects high-value testing (HVT), and awareness of potential cognitive tendencies and emotional factors (CEF). Each domain has distinct descriptors that form a behavioral anchor to aid in characterization of 'complete' diagnostic reasoning performance of learners' reasoning during a case presentation. The anchors for 'minimal' and 'partial' performance use the same descriptors that have been reworded to correspond to the extent to which the learners' presentation indicates mastery of relevant domain. The detailed descriptions of this 5-domain rubrics are given in the Supplemental Content (Figure 1).

To understand the unique contributions and psychometric properties of each descriptor in a domain to overall domain performance, they were evaluated individually. The ART, as an assessment *for* learning rubric, yields categorical, non-interval data (i.e. the minimal, partial, and complete behavioral anchors are detailed but not proportionally scaled). Therefore, the complete anchor for each domain was reconstructed to its original 3 descriptors, and a 5-point Likert-type agreement scale was applied to each descriptor to allow an interval data for the rating. Thus, the ART was reconstructed into a 15-item instrument with 3 items for each of the five domains of the ART (henceforth, the ART-reconstructed version, or ART-R) in the Supplemental content (Figure 2).

Study design

This psychometric study of the ART-R gathered validity evidence related to *internal structure* and *relationships to other variables* (Messick 1989; Cook and Beckman 2006). We investigated the three items within each domain to confirm the underlying construct of the ART-R via confirmatory factor analysis. We investigated the internal consistency reliability of the five domains of the ART-R. We evaluated the relationship between each ART-R domain score with an assessor's global assessment of diagnostic reasoning exhibited in learners' oral case presentations in an experimental module.

Experimental protocol

To investigate the 15-item ART-R, we generated representations of a broad range of learner performances. First, we generated learner presentation scripts to reflect the three levels ('complete,' 'partial,' and 'minimal') of HD, PR, PDX, and HVT and two levels ('yes' and 'no') of CEF; this resulted in 14 scripts. We then converted each of these scripts into animated video clips using the Plotagon® online software. A fractional-factorial design, generated by SAS®9.4 system, identified 18 different combinations for video clips needed to detect the domains and levels of the ART. These were then translated into the presentation scenarios by arranging the animated video clips in the suggested sequence presented in the Supplemental Content (Table 1). For instance, a scenario might contain the following sequence of clips: HD-complete, PR-partial, PDX-partial, HVT-minimal, CEF-yes. The video prototypes were evaluated by the authors and 5 clinical teacher volunteers for clarity of content and format.

Participants were academic physicians in pediatrics, internal medicine, and family medicine across all regions of the US who were recruited through a survey data collection service (SurveyHealthcare) with a $30 incentive. Respondents from other specialties or non-academic

settings were excluded through a screening question. The experimental module was hosted on the Qualtrics platform. Followed informed consent, all participants viewed a 5-minute video that introduced the concepts and terms used in the ART. This video content was derived from the original faculty training videos created by SIDM (available at https://www.improvediagnosis.org/art/). Each participant viewed two randomly selected learner presentations (out of the 18) with each presentation followed by an assessment of the diagnostic performance using a global rating scale based on overall subjective judgment (five-point rating scale; 1 = poor to 5 = excellent) followed by the ART-R.

We piloted the experimental module, monitoring response process and conducting a preliminary analysis of the first 10 completions to ascertain integrity of the data and the data collection process (i.e. logical ratings of the learner performances). There were no concerns pertaining to the quality of the data, so the data collection was continued. The study was approved by the Baylor College of Medicine institutional review board.

Assessment of validity of measurement

We conducted a confirmatory factor analysis (CFA) to validate the theoretical 5-factor structure (i.e. 5 ART-R domains). We examined the comparative fit index (CFI), standardized root mean squared residual (SRMR), and the root mean square error of approximation (RMSEA) to assess goodness of fit of this model. The CFI greater than 0.90 reflects a good model fit. SRMR less than 0.055 are ideal. The RMSEA less than 0.55 are viewed as most ideal, between 0.055 and 0.08 suggest fair model fit, values between 0.08 and 0.10 are deemed to be adequate fit, and values above 0.10 are considered a poor fit (Fabrigar et al. 1999).

Cronbach's alpha was used to determine internal consistencies within each domain. We assessed the relationships between ART-R domain scores and the global assessment of learner performance using multivariate regression analysis.

Data analysis

We performed CFA using Mplus Version 8.3 (Muthén & Muthén, Los Angeles, CA, USA). Factors were estimated with Weighted least square and variance adjusted (WLSMV) and rotated with an oblimin (Geomin) rotation that provided the best-defined factor structure. Sample size was determined using the items-to-participants criterion because we did not have a priori knowledge of communalities. The recommendation for number of participants to items (Nunnally 1978; Arrindell and van der Ende 1985; Bentler and Chou 1987) ranges from 5 to 20. Since the ART-R has 15 items, we aimed for 10 responses per item (total 150 responses).

Results

Of 237 individuals who responded to the invitation, 152 completed the study. Participants included attending physicians in family medicine (22%), internal medicine (33%), and pediatrics (46%). Sixty-one percent of participants requested more information about the ART and received the link to the tool. Since each participant assessed 2 presentations, there were 304 total assessments across the 18 scenarios.

Confirmatory factor analysis

The 5-factor model showed favorable fit indices according to CFI (0.99), RMSEA (0.097) and SRMR (0.026). The inter-factor correlations are shown in Supplemental Content (Table 2). Given high inter-factor correlations (>0.7) for HD with PR and PD with HVT, the 3- and 4-factor models were also examined for possibility of merging the factors. The statistical superiority of the 5-factor model confirmed the ART-R's 5-domain theoretical framework (Supplemental Content, Table 3). Descriptive data for all five factors is shown in Supplemental Content (Table 4). CFA generated standardized factor loading, the strength of association, of each ART-R item on the respective factors. All standardized factor loadings were well above 0.6 and statistically significant indicating that each item significantly contributes to measurement of the underlying factor and should be kept in the model (Table 1).

Internal consistency

Internal consistency of each domain was high: Cronbach's alpha was 0.90 for HD, 0.88 for PR, 0.91for PD, 0.90 for HVT, and 0.96 for CEF (Table 2). The PR domain had the lowest Cronbach's alpha (0.88) of all domains, and the PR3 (The learner's ability to employ descriptive medical terminology or semantic qualifiers in the assessment) had a lower item-total correlation (0.71) than other items. Given the favorable Cronbach's alpha for the domain, any deletion or change in wording is not needed.

Relationship to other variables

A multivariate regression analysis was performed to determine the relationship between the ART-R domains and the global assessment of the learner presentation, which is another important form of validity evidence. All five domain scores were positively and mostly significantly associated with the global assessments (Table 3). The overall model was significant, with an adjusted R^2 of 0.79. According to the Standardized Estimates, HD and PR are two strongest domains associated with global assessments. We examined collinearity diagnostics (i.e. variance inflation factor and condition index) and affirmed that the observed associations were not significantly inflated due to the correlations among the five domains.

Discussion

Building on prior evidence that demonstrated the content and response process validity of the ART (Thammasitboon et al. 2018), we derived a reconstructed version, the ART-R, using the domains and descriptors of the ART, and performed a psychometric evaluation in an experimental module. The results demonstrate validity evidence pertaining to the ART-R's internal structure and relationship to other variables. CFA confirmed the five-factor model with good fit indices consistent with the proposed theoretical framework of the ART (i.e. the five distinctive domains of

Table 1. Items included in the measurement model of the ART-R, their standardized factor loadings, standard error of loadings, *t*-statistic, and *p*-values.

Item number	Item	Standardized factor loading	Standard error	*t*-test	*p*-value
Factor 1: Hypothesis-directed information gathering (HD)					
HD1	The learner's ability to follow a clear line of inquiry towards specific diagnoses when gathering information from the patient.	0.94	0.01	70.51	<0.0001
HD2	The learner's ability to direct questions in a manner that increased/decreased the likelihood of specific diagnoses when gathering information from the patient.	0.91	0.01	68.40	<0.0001
HD3	The learner's ability to conduct the physical exam in a manner that increased/decreased the likelihood of specific diagnoses.[a]	0.87	0.01	44.93	<0.0001
Factor 2: Problem representation (PR)					
PR1	The learner's ability to give a clear synopsis of the clinical problem.	0.94	0.01	70.51	<0.0001
PR2	The learner's ability to emphasize important positive and negative findings in the assessment.	0.90	0.02	68.40	<0.0001
PR3	The learner's ability to employ descriptive medical terminology (semantic qualifiers) in the assessment.	0.84	0.02	44.93	<0.0001
Factor 3: Prioritization of the differential diagnosis (PD)					
PD1	The learner's ability to clearly rank the differential diagnoses.	0.91	0.01	66.34	<0.0001
PD2	The learner's ability to include likely and can't miss diagnoses.	0.92	0.01	79.34	<0.0001
PD3	The learner's ability to include key diagnoses in the differential diagnosis.	0.91	0.02	62.88	<0.0001
Factor 4: Diagnostic evaluation in a manner that reflects high-value testing (HVT)					
HVT1	The learner's ability to direct evaluation towards most likely and can't miss diagnoses.	0.92	0.02	61.35	<0.0001
HVT2	The learner's ability to direct evaluation in an efficient order.	0.94	0.01	78.66	<0.0001
HVT3	The learner's ability to defer tests directed towards less likely or less important diagnoses.	0.83	0.02	41.69	<0.0001
Factor 5: Awareness of potential cognitive tendencies and emotional factors (CEF)					
CEF1	The learner's ability to recognize one or more potential cognitive tendencies that might have influenced their decision.	0.95	0.01	141.62	<0.0001
CEF2	The learner's ability to recognize one or more potential emotional/situational factors that may have influenced their decision.	0.98	0.01	173.07	<0.0001
CEF3	The learner's ability to describe the ways in which cognitive/emotional/situational factors may have influenced their decision.	0.96	0.01	142.60	<0.0001

Data relate to confirmatory factor analysis (see 'Results' section). All items were retained in this model because all *p*-values were significant.
[a]Participants rated the item HD3 based on clinical information being presented by the learner and not a direct observation of the physical exam.

Table 2. Internal consistency of the ART-R: Cronbach's alpha for each ART-R domain, corrected item-total domain score correlations, and Cronbach's alpha if the item deleted from the domain.

Factor (Subscale) Items	Cronbach's alpha	Corrected item-total domain score correlation	Cronbach's alpha if the item deleted
ART-R (all fifteen items)	0.96		
Hypothesis-directed information gathering (HD)	0.90		
HD1		0.83	0.82
HD2		0.82	0.82
HD3		0.73	0.90
Problem representation (PR)	0.88		
PR1		0.80	0.81
PR2		0.80	0.80
PR3		0.71	0.88
Prioritization of the differential diagnosis (PD)	0.91		
PD1		0.81	0.88
PD2		0.84	0.86
PD3		0.83	0.87
Diagnostic evaluation in a manner that reflects high-value testing (HVT)	0.90		
HVT1		0.80	0.86
HVT2		0.82	0.84
HVT3		0.79	0.87
Awareness of potential cognitive tendencies and emotional factors (CEF)	0.96		
CEF1		0.90	0.95
CEF2		0.92	0.93
CEF3		0.92	0.94

Corrected Item-total correlation is the correlation between each item and a scale score that excludes that item. Since all values are greater than 0.4, none of the items needs to be removed. Cronbach's Alpha if item is deleted is the value of overall alpha of each domain if that item is not included in the calculations. None of the items here would substantially affect reliability if they were deleted.

diagnostic reasoning). We present evidence supporting a hypothesis of validity for the instrument based on the high internal consistency of descriptor items within a domain, acceptable inter-domain correlations, and the positive associations between the ART-R domain scores with the global assessments assigned to learners' oral presentations.

In our previous publication (Thammasitboon et al. 2018), we compared the ART with other assessment tools in the field. The theory-informed, descriptor-rich nature of the ART offers a clear structure and shared language for teachers and learners. The ART has been well received by clinical teachers given its straightforward and comprehensive

Table 3. Multivariate regression estimates (unstandardized parameter estimates, standard error, *p*-value, and standardized estimates) and collinearity diagnostics (variation inflation factor and condition index) for the effect of ART-R domains on global assessment.

Domain	Unstandardized parameter estimates	Standard error	*p*-Value	Standardized estimates	Variance inflation factor	Condition index
Hypothesis-directed information gathering	0.09	0.02	<0.0001	0.28	3.75	9.90
Problem representation	0.10	0.02	<0.0001	0.32	4.78	12.01
Prioritized differential diagnosis	0.02	0.01	0.056	0.08	2.71	14.26
High value testing	0.05	0.01	<0.0001	0.18	3.04	16.52
Metacognition	0.04	0.01	<0.0001	0.16	1.98	27.59

Multivariate regression analysis of the ART-R domains on the global assessment of the learner presentation. All five domain scores were positively and mostly significantly associated with the global assessments. The variance inflation factors (i.e. indices that measure how much the variance of an estimated regression coefficient is increased because of collinearity) are low (<10). The condition indices, another index for collinearity, are also low (<30).

domains of diagnostic reasoning based on a unified theoretical framework accompanying informative behavioral descriptors. This current study has derived an alternative form of the ART, the 15-item 5-point Likert scale ART-R, that can offer clinical teachers an additional tool for assessing learner performance. The psychometric properties presented here support the interpretations of the assessment results of the ART-R and, in turn, add validity evidence to support the original ART's domains and its behavioral descriptors.

Whereas the original ART rubric is meant to be used as a formative assessment (i.e. no stakes) to drive learning, the ART-R may be used for both 'no stakes' and 'low stakes' assessment (van der Vleuten et al. 2012; Schut et al. 2018). The ART and the ART-R can be used complementarily in various contexts and by different assessors to serve as parts of a program of assessment for competency in clinical reasoning. In programmatic assessment, each individual assessment has limited consequences (i.e. low stakes) but the aggregated assessments can be used to inform decisions about graduation and promotion (i.e. high stakes) (Schut et al. 2018). Given the assessment for learning function of the ART, the rubric may be useful during early stage of clinical training whereas the ART-R may be used later to assess and track learner progression. Additional details about the ART have been published previously (Thammasitboon et al. 2018). The accompanying the faculty training module comprising five short (3–4 minute) animation videos are available at www.improvediagnosis.org/art.

In addition to the practical use of the tool by clinical teachers, many educators and researchers desire an assessment tool with rigorous validity evidence. We propose that the ART-R provides an alternative option that is conducive for measurement of learner performance. The 15 distinctive and deconstructed steps of diagnostic reasoning with the Likert-type scale allows for discrimination of performance levels. With the robust psychometric properties, teachers and scholars can now ascertain valid and reliable interpretations of the measurements within their respective contexts.

We iteratively derived the ART-R with only three items per ART domain, so it can be used easily by clinical teachers. The distinctive domains and favorable coefficients of consistency or reliability suggest preservation of all 15 items of the ART-R. The tight correlation between the hypothesis-directed information gathering and the problem representation domains is in alignment with the theoretical framework. This also aligns with the authors' experiences where learners who gather clinical information relevant to diagnostic hypotheses are best equipped to formulate an accurate problem representation. However, CFA results still affirm that the two aspects should be kept as separate domains of assessment.

The study has several limitations. The results may not be generalizable to physicians without an interest in medical education and/or assessment. Though our participants were recruited by a survey data collection service, over half of the participants were interested in learning more about the ART. Thus, it is possible that our volunteer group was more engaged with the tool than typical faculty who might use the tool in workplace settings. We required participants to view a five-minute training video to participate in the study. The training brevity may have resulted in inadequate familiarity/comfort with the tool leading to difficulties in understanding the key concepts and discriminating minimal, partial, and complete domain performances.

The experimental setting raises questions about whether the psychometric data would be replicated in an authentic clinical environment. We did not create videos that depicted history-taking or physical examination skills so we are uncertain how using the tool in a direct observation context would impact its psychometric properties. Finally, our study lacks validity evidence regarding the consequences of the ART-R on assessors and learners; collecting these data was beyond the scope of our current research. Despite these limitations, we believe the ART-R now has validity evidence to join other tools (Norcini et al. 2003; Donato et al. 2008; van der Vleuten et al. 2008; Kogan et al 2009; Baker et al. 2015; Carter et al. 2018) that are used to assess diagnostic reasoning of learners. The ART-R has the advantage of providing the structure (via distinctive domains) and teaching language and process (via descriptors) of diagnostic reasoning to learners and teachers. It expands a versatile diagnostic reasoning assessment tool which should increase its acceptability in providing valuable formative feedback to learners throughout their clinical education.

Repeating this study in a direct observation context could provide further insights into the tool's utility. Future studies should investigate how the ART performs in its original form and function (i.e. a five-domain, behaviorally anchored rubrics to facilitate teaching and feedback) in a clinical setting. Study of the consequences of ART or ART-R implementation on teachers and learners could provide additional validity evidence for the tool (e.g. Does the learner show improvement on subsequent oral presentations?). Teachers or researchers may use either tool to track learner progression in a remediation program. Its utility should be also evaluated along with other assessment tools as a program of assessment that drive learning within competency-based education.

In conclusion, the reconstructed version of the Assessment of Reasoning Tool (ART-R) possesses validity evidence related to internal structure and relationship to other variables. We propose the use of the ART-R to guide

teaching and assessment of learners toward competency development in diagnostic reasoning.

Acknowledgements

The authors acknowledge the following members of the Society to Improve Diagnosis in Medicine Education Committee Assessment subcommittee for their many contributions to this development of the original Assessment of Reasoning Tool: Robert Trowbridge, MD, Andrew Olson, MD, Ethan Fried, MD, William Follansbee, MD, Frank Papa, DO, PhD, and Brent Smith, MD.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Funding

This research was supported by an institutional Educational Research grant from the Center for Research, Innovation and Scholarship in Medical Education, Department of Pediatrics, Texas Children's Hospital.

Notes on contributors

Satid Thammasitboon, MD, MHPE, is an Associate Professor at the Texas Children's Hospital, Baylor College of Medicine and the director for the Center for Research, Innovation and Scholarship in Medical Education.

Moushumi Sur, MD, is an Assistant Professor at the Texas Children's Hospital, Baylor College of Medicine.

Joseph J. Rencic, MD, is an Associate Professor of Medicine and the director of Clinical Reasoning Education at the Boston University School of Medicine.

Gurpreet Dhaliwal, MD, is a Professor of Medicine at the University of California San Francisco and the site director of the Internal Medicine clerkship at the San Francisco VA Medical Center.

Shelley Kumar, MS, MSc, is an Instructor/Sr. Statistician at Center for Research, Innovation and Scholarship in Medical Education, Texas Children's Hospital, Baylor College of Medicine.

Suresh Sundaram, PhD, is an Assistant Professor of Marketing, Faculty Director – Minor in Professional Selling and Sales Management, Faculty Director – Study Abroad Semester and Internships, Department of Business Administration, Alfred Lerner College of Business & Economics, University of Delaware, Newark, DE, USA.

Parthasarathy Krishnamurthy, PhD, MBA, is a Larry J. Sachnowitz Professor of Marketing, C. T. Bauer College of Business, University of Houston, Adjunct Assistant Professor of Pediatrics, Baylor College of Medicine (by courtesy), and Adjunct Assistant Professor of Anesthesiology and Pain Medicine, UTMB (by courtesy).

References

Abouna GM. 1999. The Integrated Direct Observation Clinical Encounter Examination (IDOCEE)-an objective assessment of students' clinical competence in a problem-based learning curriculum. Med Teach. 21(1):67–72.

Arrindell WA, van der Ende J. 1985. An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. Appl Psychol Measurement. 9(2):165–178.

Baker EA, Ledford CH, Fogg L, Way DP, Park YS. 2015. The IDEA assessment tool: assessing the reporting, diagnostic reasoning, and decision-making skills demonstrated in medical students' hospital admission notes. Teach Learn Med. 27(2):163–173.

Balogh EP, Miller BT, Ball JR. 2015. The path to improve diagnosis and reduce diagnostic error. In: Ball JR, Balogh EP, Miller BT, editors. Improving diagnosis in health care. Washington (DC): National Academies Press; p. 355–402.

Bentler PM, Chou CH. 1987. Practical issues in structural modeling. Sociol Methods Res. 16(1):78–117.

Bordage G. 2007. Prototypes and semantic qualifiers: from past to present. Med Educ. 41(12):1117–1121.

Carter C, Akar-Ghibril N, Sestokas J, Dixon G, Bradford W, Ottolini M. 2018. Problem representation, background evidence, analysis, recommendation: an oral case presentation tool to promote diagnostic reasoning. Acad Pediatr. 18(2):228–230.

Cook DA, Beckman TJ. 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 119(2):166.e7–166.e16.

Custers EJ. 2015. Thirty years of illness scripts: theoretical origins and practical applications. Med Teach. 37(5):457–462.

Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, Ratcliffe T, Gordon D, Heist B, Lubarsky S, et al. 2019. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med. 94(6):902–912.

Donato AA, Pangaro L, Smith C, Rencic J, Diaz Y, Mensinger J, Holmboe E. 2008. Evaluation of a novel assessment form for observing medical residents: a randomised, controlled trial. Med Educ. 42(12):1234–1242.

Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. 1999. Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods. 4(3):272–299.

Gruppen LD. 2017. Clinical reasoning: Defining it, teaching it, assessing it, studying it. West J Emerg Med. 18(1):4–7.

Kogan JR, Holmboe ES, Hauer KE. 2009. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA. 302(12):1316–1326.

Messick S. 1989. Validity. In: Linn RL, editor. Educational measurement. 3rd ed. New York (NY): Macmillan; p. 13–104.

Norcini JJ, Blank LL, Duffy FD, Fortna GS. 2003. The mini-CEX: a method for assessing clinical skills. Ann Intern Med. 138(6):476–481.

Nunnally JC. 1978. Psychometric theory. 2nd ed. New York (NY): McGraw-Hill.

Schut S, Driessen E, van Tartwijk J, van der Vleuten C, Heeneman S. 2018. Stakes in the eye of the beholder: an international study of learners' perceptions within programmatic assessment. Med Educ. 52(6):654–663.

Thammasitboon S, Rencic JJ, Trowbridge RL, Olson AP, Sur M, Dhaliwal G. 2018. The assessment of reasoning tool (ART): structuring the conversation between teachers and learners. Diagnosis. 5(4):197–203.

van der Vleuten CPM, Norman GR, Schuwirth LWT. 2008. Assessing clinical reasoning. In: Higgs J, Jones MA, Loftus S, Christensen N, editors. Clinical reasoning in the health professions. 3rd ed. Edinburgh: Elsevier; p. 21–55.

van Der Vleuten CPM, Schuwirth L, Driessen E, Dijkstra J, Tigelaar D, Baartman L, van Tartwijk J. 2012. A model for programmatic assessment fit for purpose. Med Teach. 34(3):205–214.

Young M, Thomas A, Lubarsky S, Ballard T, Gordon D, Gruppen LD, Holmboe E, Ratcliffe T, Rencic J, Schuwirth L, et al. 2018. Drawing boundaries: the difficulty in defining clinical reasoning. Acad Med. 93(7):990–995.