Does it make sense to use written instruments to assess communication skills? Systematic review on the concurrent and predictive value of written assessment for performance

Authors

Claudia Kiessling[1], Noelle Junod Perron[2], Marc van Nuland[3], Maria Magdalena Bujnowska-Fedak[4], Geurt Essers[5], Ragnar M. Joakimsen[6], Peter Pype[7], Zoi Tsimtsiou[8]

Affiliations

[1]Chair for the Education of Personal and Interpersonal Competencies in Health Care, Witten/Herdecke University, Witten, Germany

[2]Unit of Development and Research in Medical Education and Department of community health and medicine, Geneva Faculty of Medicine and Medical Directorate, Geneva University Hospitals, Geneva, Switzerland

[3]Academic Center for General Practice, Leuven University, Leuven, Belgium

[4]Department of Family Medicine, Wroclaw Medical University, Wroclaw, Poland

[5]Network of GP Training Programs in the Netherlands, Utrecht, The Netherlands

[6]Department of Clinical Medicine, Faculty of Health Sciences, UIT The Artic University of Norway and Department of Internal Medicine, University Hospital of North Norway, Tromsø, Norway

[7]Department of Public Health and Primary Care, Ghent University, Ghent, Belgium

[8]Department of Hygiene, Social-Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

Corresponding author at

Prof. Dr. med. Claudia Kiessling, MPH

Chair for the Education of Personal and Interpersonal Competences in Health Care, Witten/Herdecke University, Witten, Faculty of Health

Alfred-Herrhausen-Str. 50

58455 Witten, Germany

Tel: +49(0)2302 926-477

E-Mail: claudia.kiessling@uni-wh.de

Key words

Communication skills; written assessment, performance-based assessment, health professionals

Highlights

Little is known about the possible association between learners' results in written and performance-based assessments of communication skills (CS) either in concurrent or predictive study designs. The correlation between learners' scores in written and performance-based assessments were found to be low to medium.

Written assessments are limited performance predictors and cannot replace performance-based assessments.

Within longitudinal assessment programs, triangulation of assessment instruments, including written and performance-based assessment, is recommended.

Systematic reporting of written assessment instruments' including psychometric properties is essential to improve the interpretation of future findings and could contribute to improved research regarding predictive validity for performance.

Article: 4089 words

Abstract

Objectives: To evaluate possible associations between learners' results in written and performance-based assessments of communication skills (CS), either in concurrent or predictive study designs.

Methods: Search included four databases for peer-reviewed studies containing both written and performance-based CS assessment. Eleven studies met the inclusion criteria.

Results: Included studies predominantly assessed undergraduate medical students. Studies reported mainly low to medium correlations between written and performance-based assessment results (Objective Structured Clinical Examinations or encounters with simulated patients), and gave correlation coefficients ranging from 0.13 to 0.53 ($p<0.05$). Higher correlations were reported when specific CS, like motivational interviewing were assessed. Only a few studies gave sufficient reliability indicators of both assessment formats.

Conclusions: Written assessment scores seem to predict performance-based assessments to a limited extent but cannot replace them entirely. Reporting of assessment instruments' psychometric properties is essential to improve the interpretation of future findings and could possibly affect their predictive validity for performance.

Practice implications: Within longitudinal CS assessment programs, triangulation of assessment including written assessment is recommended, taking into consideration possible limitations. Written assessments with feedback can help students and trainers to elaborate on procedural knowledge as a strong support for the acquisition and transfer of CS to different contexts.

1 Introduction

Health professional learners spend most of their time acquiring scientific knowledge, especially in the early years of training [1, 2]. Despite the current trend to base curricula and training on competencies, there is no doubt that scientific knowledge is necessary within health professional education and practice [3, 4, 5, 6]. Medical doctors and other healthcare professionals need a large amount of basic knowledge for clinical reasoning, diagnosing, decision-making, and treating patients [7, 8, 9]. However, when it comes to good clinical communication skills (CS), the role of knowledge within the formation of these skills is debatable. CS training is mainly experiential and reflective- with role-plays, simulations, and sessions on values and attitudes- and not so much based on the acquisition of knowledge [10, 11, 12, 13, 14, 15, 16].

Whereas clinical reasoning, diagnosing, decision making, etc. are viewed as cognitive skills with a clear fundamental knowledge [7, 8, 9, 17, 18], definitions of what constitutes good clinical communication seems to vary. Is it a cognitive skill, a social skill, emotional intelligence, a personality trait, an attitude, or a competency? As examples, Weinert defined competency as 'cognitive abilities and skills available to individuals to solve specific problems, as well as the associated motivational, volitional, and social readiness and ability to successfully use problem-solving in various situations' [19, 20]. According to Hargie, individuals apply in social interactions a set of purposeful, interconnected, situationally appropriate social behaviors that are learned and controlled [21, 22]. Social skills include verbal and non-verbal behaviors and are influenced by contextual factors. Emotions and personal involvement play an important role as well as reflecting on the process. These definitions include knowledge about processes, conditions, and contexts. It also includes the cognitive ability to perceive, apply, organise, and control. In the words of Krathwohl [23] who revised Blooms' taxonomy of educational objectives for the cognitive domain [24], it contains factual knowledge, conceptual knowledge, procedural knowledge, and metacognitive knowledge. What does this mean for the assessment of CS?

A well-known framework to categorize assessment approaches is the Miller's pyramid of the assessment of clinical skills/competence/performance [25]. The base of the Miller's pyramid is "some assurance that a student, a resident, a physician knows what is required" to carry out professional activities effectively [25, p.S63]. Written instruments are commonly used to test this factual knowledge. At the next level, learners must *know how* to use that knowledge appropriately. More complex processes are involved including procedural, conceptual, and metacognitive knowledge [23]. Miller recommends written instruments including patient vignettes to address this type of applied context-based knowledge [25]. At the *shows how* level, a learner is able to show a performance, e.g. in a simulated situation, whereas the *does* level relates to how he or she will act in everyday practice. To assess the *shows how* level, simulated encounters with standardized patients (SP) e.g. an

Objective Structured Clinical **E**xamination (OSCE) are most commonly used. Workplace-based assessment (WBA) is recommended for the *does* level: the action in real-life situations.

When it comes to recommendations about how to assess CS, the role of written assessments has either been described as being of limited use [26] or not included into recommendations at all [27, 28]. Two guidelines on the assessment of clinical communication explicitly refer to the Miller pyramid [29, 30]. Both guidelines acknowledge the importance of assessing all components of clinical competence "to fully appraise the abilities of an individual" [29, p.5]. However, Laidlaw et al. state, that "there is surprisingly little evidence that shows that grades on the different levels of the skills pyramid correlate with each other" [29, p.5]. The focus of the assessment of practical skills in these well-recognized guidelines and publications might have led to the impression that assessment of knowledge plays a minor role in the field of communication.

However, if we follow the idea that clinical communication is a competency or social skill, which includes cognitive abilities [31, 32] and can be learnt, then the acquisition of knowledge about clinical communication would have a predictive role on the performance of effective clinical communication behaviour. However, data are scarce about associations between written assessment and performance-based assessment in the field of CS trainings. Predictive validity of written assessments for performance might support the usefulness of written assessment in CS training and help answer the question whether it makes sense to use written instruments for the assessment of CS.

This systematic review is an extension of a recently published scoping review about written assessment of CS [33]. The aim of the scoping review was to investigate the extent, range and nature of published research activity regarding the use of written assessments of the cognitive component of communication skills in health professionals' education. A quality assessment of studies was not intended due to the heterogeneity of studies and lack of psychometric data in the majority of included articles. However, a few articles provided sufficient information about both psychometrics and associations between written assessments and performance-based assessments. The aim of this systematic review was to explore the available body of evidence in the literature about the possible association between learners' results in written assessment (knows and knows how) and performance-based assessment (shows and does)."

2 Methods

Data sources

We conducted a formal literature search using the Preferred Reporting Items for Systematic Review and Meta-analysis Protocols (PRISMA_P) [34].

Research question

Do learners' results in written assessments of CS correlate with their results in performance-based assessments either in concurrent or predictive study designs?

Search strategy and eligibility criteria

As mentioned before, this systematic review is an extension of a recently published scoping review about written assessment of CS [33]. The search strategy used four key terms: 1) communication, 2) educational measurement/written assessment, 3) computer simulation/computer/video recording/video-based test measurement/paper-based, 4) knowledge/cognition/clinical competence/skill. We searched Pubmed, Embase, Cinahl and Psychinfo. Details of our search strategy are given in Appendix 1.

We used the same eligibility criteria as for the scoping review [33]. Inclusion criteria were studies published from January 1995 until September 2021, written in English or any other language spoken by the research team (Dutch, French, German, Greek, Norwegian, Polish, and Portuguese).Studies of interest focused on learners in the health professions at the undergraduate, postgraduate or continuous training levels, reporting an educational assessment (test, exam, measurement) and using or aiming at using empirical data to measure the cognitive abilities of individual learners in the field of clinical communication (patient provider communication). Study designs included protocols, development, usage and/or validation of a written assessment, interventions, correlation studies, pre and/or post interventions or those which compared different types of interventions. Using a snowball approach, we added articles on the reference lists if they met the inclusion criteria mentioned above and were not listed in the initial search. Exclusion criteria were studies focusing on patients, pupils, simulated patients or teachers, using designs such as needs assessments, surveys on patient satisfaction, reporting exclusively outcome measures such as the performance of skills, behaviors, attitudes, and self-reported perceptions. Studies assessing knowledge regarding intra- or interprofessional team communication were excluded since communication with patients is only a small part of interprofessional communication. We did not include grey literature.

For the systematic review, we enclosed an additional inclusion criterion, namely studies in which empirical data of learners' results in both written and performance-based assessment as well as associated measures between those results that were reported. Studies without empirical data or reporting single measurement outcomes were excluded as well as studies assessing knowledge and behavior/performance regarding team communication. We also did not include grey literature. The search strategy and selection process were conducted in two stages: initially for the scoping review between 2018 and 2020 and an update in September 2021.

Study selection

We used EndNote to collect all the references and imported all titles into Rayyan software [35]. Initial duplicates were removed, and titles and abstracts were screened for inclusion. Selection criteria for the update were the same as for the initial scoping review. We had gained a considerable amount of experience and reliability in reviewing the titles and abstracts for the scoping review in pairs (> 20.000 screened titles divided among four pairs) with good interrater agreement, while the reading and analysis of titles/abstracts published between 2020 and 2021 were conducted by single investigators (600 titles per investigator). Studies categorized as "include" and "unsure" were re-analyzed by two investigators. In case of disagreement, the study was categorized as "include". Where full texts were not available, authors were contacted requesting the full text. For two articles, we did not receive a reply. The flowchart of the procedure that was followed is given in Figure 1.

Data extraction and quality assessment

Two investigators (CK, ZT) extracted data from the full text articles coming from the scoping review (January 1995 until July 2020) and from the updated search (July 2020 until September 2021) using an adapted version of the excel sheet created for the scoping review [33]. Quality of studies was assessed according to Hurwiler et al. [37] and Hammick et al. [38] by evaluating the methodological quality of the study as well as the quality of the information provided in the article. We also assessed the strength of findings on a scale from 1 to 5 (see Table 1 for the detail of scores). Two researchers (NJP, ZT) independently evaluated the included articles in parallel. Differences of perceptions were resolved by discussion. A third author (CK) was involved in case of disagreement.

Data synthesis

We were not able to identify a sufficiently large and homogeneous group of studies to permit quantitative synthesis. Hence, we synthesized the results narratively (Table 1). Selection and extraction procedures and data synthesis were discussed among the whole researcher group in regular online meetings.

3 Results

We found 11 articles reporting on studies that included both written assessment tests and assessment of performance in CS [39-49] (Table 1). As two articles described the same written assessment instrument and built on each other, both were summarized in Table 1 [42, 43].

Description of setting and population of the studies

Of the eleven studies, three were performed in the US [39, 40, 41], two in the UK [42, 43], two in Germany [44, 45], one in the Netherlands [46], one in Belgium [47], one in Israel [48], and one in Malaysia [49], published between 2000 and 2021. The population of the studies included

undergraduate medical students [42-49], residents [40], staff clinicians [41], and counsellors in substance abuse community programs [39].

Characteristics of written communication instruments

Written communication tests measured general CS (e.g. knowledge about relationship building or providing structure) [40, 41, 43, 46, 47, 49], specific CS- like motivational interviewing [39, 41], breaking bad news [48] - or both [44, 45].

The written communication tests used were: Objective Structured Video Examination (OSVE) [42, 43], Helpful Responses Questionnaire (HRQ) [39], Helpful Responses Questionnaire- contingency management (HRQ-CM) [41], Communication Skills Video Assessment (CSVA) [49], Reflection Evaluation for Learners' Enhanced Competencies (REFLECT) with additional communication parameters [48], Computer-Based Test including a Situational Judgement Test (CBT) [44], Knowledge test about Communication Skills (KCS) [46], three Situational Judgment Tests (SJT) [40, 45, 47], including Video-based Single-Choice Examination (VSE) [45].

All tests used-patient vignettes, either in the form of videos [40, 41, 43, 45, 47], or with written scenarios [37, 38, 39, 42, 44] as stimuli. Prompts or lead-in questions were designed to activate either knowledge-oriented cognitive processes (e.g. 'explain', 'analyze', 'evaluate') [38, 42, 44, 45], behavior-oriented cognitive processes (e.g. 'apply', 'create", 'what would you do') [37, 39, 46] or both [40, 41, 43, 47]. One study used written prompts for reflective writing of personal patient contact [47]. Expected candidate response format varied with selected response (e.g. multiple choice questions) [38, 42, 43, 44, 45] or constructed response formats (e.g. short answer questions) [37, 39, 40, 41, 46, 47].

Of the eleven studies, eight studies reported indicators for reliability either test reliability, in terms of reporting internal consistency [40, 41, 44, 45, 46, 47] measured by Cronbach's alpha (ranging from α=.55 [45] to α=.86 [40]) or rater reliability for coding learners' responses [39, 42] calculated mostly by intraclass correlation (ranging from ICC=.79 [39] to ICC=.96 [41]).

Characteristics of performance assessments of communication skills

For the assessment of performance in CS, OSCEs were used in four studies (one to five stations per study) [42, 43, 45, 49], rated SP encounters in four studies (one to four encounters per study, live or videotaped) [41, 44, 46, 48], workplace-based assessment related to the Accreditation Council for Graduate Medical Education (ACGME) and assessment by a faculty committee [38], assessment of randomly selected audiotaped segments of counselling sessions [39], and internship and job performance ratings [47]. The number of OSCE stations and SP encounters ranged from one [41, 48, 49] to five [45] stations or encounters **with an observation time between five minutes [39] and 60 minutes [46]**.

The performance was rated by expert examiners/raters in all eleven studies (one to two raters per study). In two studies [42, 43], SP also rated participants' performance. The following performance-based assessment tools were used: Liverpool Communication Skills Assessment Scale - LCSAS [42, 43], MITI 2.0 [39], specific forms used for communication skills assessment by each institution [45, 49], Berlin Global Rating (BGR) [44], Observing Patient Involvement Instrument (OPTION) [44], MAAS-Global [46], Contingency Management Competence Scale (CMCS) [41], ACGME milestone performance, Professionalism Multisource Assessment (MPA) and USMLE scores [40], internship and job performance rating score sheets [47], Breaking Bad News Assessment Schedule (BAS) [48] and SPIKES Questionnaire [48].

Reliability of performance measures was calculated using Cronbach's alpha in five studies [40, 44, 45, 46, 48] ranging from .49 [46] to .92 [40]. Two studies reported Intraclass Correlations on rater agreement [39, 41].

Correlation of written test results with performance assessment

Nearly all studies reported positive low to moderate [50] statistically significant correlation coefficients between written and performance assessments ranging from r= 0.13 [40] to 0.53 [41]. Of the eleven studies, six studies referred to concurrent validity by reporting low correlation coefficients (.10 - .29) between scores collected at one survey time point [40**,** 42, 43, 44, 46, 48]. Three studies reported low correlations in terms of predictive validity **(.13 - .28)** because performance was measured from three months to up to nine years later than when the written test was administered [40, 47, 49]. Moderate correlations (.30 - .49) were reported by three studies [39, 46, 48], one of these also reporting predictive validity [39]. Only one study reported ~~a~~ strong correlations between written assessment and performance-based assessment directly after the intervention and three months later (r= .53; r = .52) [41].

A few authors also reported non-significant low correlation coefficients, e.g. with single performance measures [40, 44], single study locations [46], subscales of performance measures [39] or performance ratings used by simulated patients [42, 43]. One author team reported a significant negative low correlation between an OSVE results and an OSCE result 17 months after the OSVE was administered [42, 43].

Quality Ratings of Included Studies

As two studies used the same written instrument and built upon each other, quality assessment for both studies were aggregated. Four studies achieved a high methodological quality score ("4" on a 1-5 Likert scale) [39, 42/43, 44, 48], while the six remaining studies were rated as moderate ("3" on a 1-5 Likert scale) [40, 41, 45, 46, 27, 49]. Eight studies respectively achieved a high score ("4") on the quality of the information provided in the article, [39, 40, 41, 42/43, 44, 45, 47, 48], while the other

two obtained a moderate one (“3”) [46, 49]. Finally, four studies achieved a high (“4”) on the strength of findings [39, 42/43, 44, 48] and six a moderate quality score (“3”) [40, 41, 45, 46, 47, 49]. The most common threats to quality were the small numbers of participants, the lack of information about participants’ sampling, the low internal consistency or inter-station reliability and the missing information on the reliability and validity of some written tests and performance assessment methods used.

The characteristics of each study - including first author’s name, date of publication, the country where the study took place, the description of the participants, the written assessment instrument(s) used, the way the performance in communication skills was assessed, and the correlation coefficients are presented in detail in Table 1.

4 Discussion and Conclusion

4.1 Discussion

Our review has synthesized the evidence about the effectiveness of learners’ results in written assessment in predicting their results in performance-based assessment. We extracted eleven studies reporting on correlation coefficients between written (*knows* and *knows how* level of Miller’s pyramid) and performance-based assessments (*shows* and *does* level of Miller’s pyramid). We were interested in whether there is an evidence base to support the use of written instruments for the assessment of communication skills. By finding meaningful associations and predictive validity for the performance of CS, the question can be answered with “yes”.

Who, what and how was assessed?

Study population predominately comprised undergraduate medical students. Interestingly, all written assessment instruments used patient vignettes as **stimuli**, indicating that these assessment instruments focused on higher order cognitive processes such as analysis, application, creation, and reflection (*knows how* level of Miller’s pyramid). In other words, the assessment of factual knowledge played a minor role. This is in line with general recommendations of written assessment. Schuwirth et al. examined case-based questions versus factual knowledge questions in the field of clinical expertise and found that **former** questions lead to higher thinking processes (problem-solving ability) than those elicited by factual knowledge questions. [51]

Regarding the assessment of performance, OSCEs and videotaped SP encounters were predominantly employed. Only few studies measured performance in the workplace [39, 40, 47].

OSCEs were quite short and the number of videotaped patient and SP encounters to measure performance were limited to one or two which might be a reason for low reliability and therefore limited generalizability of their results. Brannick et al. [52] showed in their meta-analysis that several factors influenced reliability such as the number of stations and total testing time. Thus, OSCEs in general yield a wide variation in reliability scores. **Van Nuland et al. reported that OSCEs focusing specifically on communication skills needed between 6 and 18 cases with observation times of 60 to 360 minutes to reach acceptable reliability [53]. In a study published by Newble and Swanson, the recommended testing time for the reliable assessment of clinical competence in OSCEs was at least four hours [54]. Reported numbers of videotaped patient or SP encounters used for assessment purposes range from two to eight in published studies [55, 56, 57, 58, 59, 60, 61].** Regarding the methodical quality of the studies, the results revealed a heterogeneous picture.

Reported associations between written and performance-based assessments

Overall, low to medium statistically significant correlations were found between learners' results in written and performance-based assessment. Several authors have described correlation coefficients between written and performance assessment scores in other clinical domains like general clinical skills or competencies and clinical reasoning. In the Federal Licensing Examination in Switzerland for example, Berendonk et al found correlations between multiple choice examination and CS scores in clinical skills examinations of r=.39 **[62]**. Brailovsky et al. reported correlations between a script concordance test and OSCE in the field of clinical reasoning of r=.34 and concluded that a written test was able to predict performance to a small amount **[63]**. Remmen et al. reported that a written test of skills was able to predict OSCE scores, ranging between.35 to .48. However, he also stated that a written test would be unable to replace the OSCE for the assessment of individual learners **[64]**. Ram et al. reported the predictive value of medical knowledge tests for actual job performance from .43 to .56, which was comparable to the predictive value of OSCE scores **[65]**.

These reported correlation coefficients in the field of general clinical competencies and clinical reasoning were within the range of the correlation coefficients that we found for communication skills, whereas a direct comparison is difficult due to the different types of correlation coefficients reported and variance in study design and purpose. We found highest correlation coefficients within studies assessing challenging communication like motivational interviewing and breaking bad news. A communality could be that they all assessed rather focused and small constructs. Other authors mainly assessed general communications or a mixture of general CS with specific communication challenges. Therefore, one conclusion is that correlation coefficients become higher when the theoretical construct to be assessed is focused and written and performance assessment measure the same theoretical construct. The type of stimulus (e.g. written or video-based patient vignettes) and response format (e.g. selected or constructed response formats) did not seem to have an

influence on the strength of the correlation. This result is in line with previous research showing that validity – what is being measured – is not so much determined by the response format as by the stimulus format **[66]**. Concerning the stimulus, Lievens & Sackett demonstrated that a video-based version of an interpersonally oriented SJT had higher predictive and incremental validity for predicting interpersonally oriented criteria than did the written version. However, further research is indicated to explore the influence of different stimulus formats of written assessment on performance-based assessment **[67]**.

Correlation coefficients seem to be stable over time because we found low but significant correlations in long time surveys from one year later up to nine years [40, 47, 49]. Number of studies with information about predictive validity is very small and one study reported negative correlations [43]. However, our results as well those of prior studies do not allow us to answer the question "how much correlation is enough to be able to predict behavior on the basis of written assessment results".

Further suggestions to increase the strength of correlation between written and performance-based assessment and therefore improve the potential to predict behaviour based on written test results is to focus the assessments on a well described construct (e.g. breaking bad news, shared decision making, motivating, etc.), to improve the reliability of instruments either by increasing the number of items and stations [52, **68**] or by highlighting the aspect of transfer between different tasks **[69]**.

Improving Reliability

Out of the six studies reporting on Cronbach's alpha for written assessment instruments, only two showed sufficient reliability [40, 41]. Regarding performance-based assessments, interstation reliability was reported in five studies, while only three of them showed sufficient reliability coefficients [40, 44, 48]. One major quality indicator of assessment instruments is internal consistency or inter-item/station reliability mainly calculated by Cronbach's alpha. According to Downing, the most frequently asked question about reliability is "how much reliability is enough?" **[70]**. For examinations like end-of-year summative examinations, one would expect reliability to be in the range of .80–.89. For assessments like formative or summative classroom-type assessments, one might expect reliability to be in the range of .70–.79 [68]. Brannick et al. report in their systematic review of the reliability of OSCE scores that alpha across stations was .66. [52].

Low reliability could imply a higher level of error, which means that due to low reliability, the likelihood to establish meaningful correlations between different instruments decreases. The ability to generalize from a test result to a broader domain depends on the degree to which the examinee's performances on selected assessment tasks are consistent. Then, generalizable conclusions seem appropriate. If the performances were not consistent, then such generalizations are not appropriate and associations with other measurements become less likely **[71]**. We fully support the

recommendation to improve reliability by increasing the number of items or stations [52, **68**]. An additional approach could be to think about why we see high variance in task performance and how a good knowledge base could help to decrease this variance and improve the ability to transfer one skill from one task to another.

Does it make sense to use written assessment? How can a good knowledge base improve the application of skills in varied contexts?

Transfer has been defined as "the ability to utilize a well understood problem to provide insight and structure for a less understood problem" [**72** p. 294, quoted in **73**]. It is related to meaningful learning and contextual application by understanding and making sense of new knowledge and skills [63]. Higher order skills like CS are context- and content-bound [**69, 74, 75**]. Studies have shown that a knowledge base helps learners in these transfer processes of skills. According to Perkins [**69**], structural knowledge is a prerequisite to valid performance assessment scores because structural knowledge leads to better transfer. An example of how knowledge informs the acquisition of skills is the four-step approach to acquire practical skills according to Peyton, which includes the steps of demonstration, deconstruction, comprehension, and performance [**76**]. The approach explicitly includes procedural knowledge into the learning process to support the learner's comprehension and performance of skills. Trainers need procedural knowledge to explain the single steps of applying a skill. Although studies are lacking about the effectiveness of such model for CS training, one can consider that emphasizing first the cognitive dimensions of communication before practising CS may improve skills acquisition, since it is clear that CS require knowledge about processes and contexts and conditions. Benefits and value in utilizing written assessment accrue for students and trainers in providing feedback and enabling them to deconstruct cognition about skills, and elaborate on procedural knowledge. Both provide a strong support for the acquisition and transfer of skills to different contexts. As cognition is associated with performance, there is a need to include conceptional and procedural knowledge into CS training. Knowledge can support the transfer of skills. Teachers who can deconstruct their knowledge to explain their use of skills on specific contexts help students elaborate their communicative competence e.g. by worked examples. Knowledgeable students will be the knowledgeable teachers of the future.

Strengths and limitations of this review

This is the first review that systematically investigates the associations between written and performance-based assessment in CS and might therefore be useful in understanding what to expect when using written assessment and enabling teachers to use it more effectively. Additionally, it could be a valuable starting point for more focused research in this field. However, some limitations may affect the interpretation of the findings. Firstly, the number of the studies included was small, since the association between written and performance-based assessment is an understudied topic.

Moreover, most of these studies examined undergraduate medical students, and therefore other levels of education and other professional contexts are under-represented, since they have not as yet been studied. Furthermore, the quality assessment of the included studies revealed a large heterogeneity of study designs and missing information e.g. confidence intervals, that complicated the appraisal of the quality of the information presented in a few cases and prevented us from conducting any further quantitative analysis. Finally, the low reliability coefficients for instruments and performance-based assessments in some studies might also have an influence on the explanatory power of this review.

4.2 Conclusion

While written assessment scores seem to be able to predict performance-based assessments to a small degree validating the use written instruments for the assessment of CS, they are limited substitute for OSCEs or workplace-based assessments

In terms of future research implications and conclusions, we would emphasize the importance of reporting the psychometric properties of instruments and the validation process of written assessment tools, providing sufficient information about association measurements. This would allow a better interpretation of the findings presented in the studies. Based on what we found, even a low, but significant correlation seems to be indicative of the tool's validity. Finally, further research is needed with predictive studies in a broad variety of health professional trainees including workplace-assessments in order to provide a better insight into the optimal use of written assessment in CS.

4.3 Practice Implications

Written assessment can be used for the triangulation of assessment tools in longitudinal assessment programs [**77**] to create the full picture of learners' achievement and progress in training. In an early phase of learning, formative written assessment might help to detect students with deficits who needs further support and remediation [**78**] because written assessments can be employed for large student groups and are not as time- and resource intense as OSCEs [**79**]. Further research is indicated to investigate the use of written assessment for these purposes.

Acknowledgement

We gratefully thank the librarians from the Geneva, Ghent and Tromso Universities for their help in the literature search. We also thank Stephanie Matthews and Anne Marie Regan for thoroughly proofreading the manuscript.

References

[1] Cooke M, Irby DM, Sullivan W, Ludmerer KM. American Medical Education 100 Years after the Flexner Report. N Engl J Med 2006;355:1339-44.

[2] Dickinson BL, Gibson K, Van Der Kolk K, Greene J, Rosu CA, Navedo DD, Porter-Stransky KA, Graves LE. "It is this very knowledge that makes us doctors": an applied thematic analysis of how medical students perceive the relevance of biomedical science knowledge to clinical medicine. BMC Med Educ 2020;20:356.

[3] Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting Paradigms: From Flexner to Competencies. Acad. Med 2002;77:361-7.

[4] Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teacher 2007; 29:642-7.

[5] Swing SR. The ACGME outcome project: retrospective and prospective, Med Teacher 2007; 29:648-54.

[6] Whitehead C. Scientist or science-stuffed? Discourses of science in North American medical education. Med Educ 2013; 47:26–32.

[7] Boshuizen HPA, Schmidt HG. On the Role of Biomedical Knowledge in Clinical Reasoning by Experts, Intermediates and Novices. Cognitive Science 1992; 16:153-84.

[8] Bordage G. Elaborated knowledge: a key to successful diagnostic thinking. Acad Med 1994; 69:883-5

[9] Charlin B, Tardif J, Boshuizen HPA. Scripts and Medical Diagnostic Knowledge: Theory and Applications for Clinical Reasoning Instruction and Research. Acad Med 2000; 75:182–90.

[10] Chant S, Jenkinson T, Randle J, Russell G, Webb C. Communication skills training in healthcare: a review of the literature. Nurse EducToday 2002; 22:189-202.

[11] Smith S, Hanson JL, Tewksbury LR, Christy C, Talib NJ, Harris MA, Beck GL, Wolf FM. Teaching Patient Communication Skills to Medical Students : A Review of Randomized Controlled Trials. Eval Health Prof 2007; 30: 3-21.

[12] Lane C, Rollnick S. The use of simulated patients and role-play in communication skills training: A review of the literature to August 2005. Pat Educ Couns 2007 67:13–20.

[13] Hargie O, Boohan M, McCoy M, Muprhy P. Current trends in communication skills training in UK schools of medicine. Med Teacher 2010; 32:385-91.

[14] Berkhof M, van Rijssen J, Schellart AJM, Anema JR, van der Beek AJ. Effective training strategies for teaching communication skills to physicians: An overview of systematic reviews. Pat Educ Couns 2011; 84:152–62.

[15] Gutiérrez-Puertas L, Márquez-Hernández VV, Gutiérrez-Puertas V, Granados-Gámez G, Aguilera-Manrique G. Educational Interventions for Nursing Students to Develop Communication Skills with Patients: A Systematic Review. Int. J. Environ. Res. Public Health 2020; 17:2241.

[16] Stiefel F, Barth J, Bensing J, Fallowfield L, Jost L, Razavi D, Kiss A. Communication skills training in oncology: a position paper based on a consensus meeting among European experts in 2009. Annals of Oncology 2010; 21:204-7.

[17] Van Lehn K. Cognitive skill acquisition. Annu Rev Psychol 1996. 47:513–39.

[18] Charlin B, Lubarsky S, Millette B, Crevier F, Audetat MC, Charbonneau A, Fon NC, Hoff L, Bourdy C. Clinical reasoning processes: unravelling complexity through representation. Med Educ 2012: 46: 454-63.

[19] Weinert FE. Concept of competence: A conceptual clarification. In: Rychen DS, Salganik LH (Eds). Defining and selecting key competencies. Seattle: Hogrefe & Huber; 2001; pp. 45–65.

[20] Weinert FE. Concepts of Competence. Contribution within the OECD Project Definition and Selection of Competencies: Theoretical and Conceptual Foundations (DeSeCo). Neuchatel: Bundesamt für Statistik; 1999; pp. 3-34.

[21] Hargie O. Skill in theory: Communication as skilled performance. In: Hargie O, Ed. The handbook of communication skills. London: Routledge; 2006; p.7-36.

[22] Hargie O. Skill in practice. An operational model of communicative performance. In: Hargie O, Ed. The handbook of communication skills. London: Routledge; 2006; p.37-69.

[23] Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory into Practice 2002, 41:12-18.

[24] Bloom BS. Taxonomy of educational objectives. New York: McKay; 1956.

[25] Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990; 65: S63-S67.

[26] Duffy, F.D., Gordon, G.H., Whelan, G., Cole-Kelly, K., Frankel, R., Buffone, N., Lofton, S., Wallace, M., Goode, L. and Langdon, L., Assessing competence in communication and interpersonal skills: the Kalamazoo II report, Acad Med 2004; 79:495-507.

[27] Schirmer JM, Mauksch L, Lang F, Marvel MK, Zoppi K, Epstein RM, Brock D, Pryzbylski M. Assessing Communication Competence: A Review of Current Tools. Family Med 2005; 37:184-92.

[28] Boon H, Stewart M. Patient-physician communication assessment instruments: 1986 to 1996 in review. Pat Educ Couns 1998; 35:161-76.

[29] Laidlaw A, Hart J. Communication skills: an essential component of medical curricula. Part I: assessment of clinical communication. AMEE guide 51. Dundee: AMEE. 2011.

[30] Kiessling C, Tsimtsiou Z, Essers G, van Nuland M, Anvik T, Bujnowska-Fedak MM, Hovey R, Joakimsen R, Junod Perron N, Rosenbaum M, Silverman J. General principles to consider when designing a clinical communication assessment program. Pat Educ Couns 2017; 100:1762-68.

[31] Kinderman P, Humphris G. Clinical communication skills teaching: the role of cognitive schemata. Med Educ 1995; 29:436-42.

[32] Hulsman RL, Mollema ED, Hoos AM, de Haes JC, Donnison-Speijer JD. Assessment of medical communication skills by computer: assessment method and student experiences, Med Educ 2004; 38:813-24.

[33] Junod Perron N, Pype P, van Nuland M, Bujnowska-Fedak MM, Dohms M, Essers G, Joakimsen R, Tsimtsiou Z, Kiessling C. What do we know about written assessment of health professionals' communication skills? A scoping review. Pat Educ Couns 2022; 105:1188-1200.

[34] Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015;4:1.

[35] Rayyan QCRI. https://rayyan.qcri.org/ (Accessed Feb, 20th 2022).

[36] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.

[37] Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. JAMA. 2002;287(21):2801–4.

[38] Hammick M, Dornan T, Steinert Y. Conducting a best evidence systematic review. Part 1: from idea to data coding. BEME Guide No. 13. Med Teach. 2010;32(1):3–15.

[39] Smith JL, Bertone P, Carpenter KM, Wain RM, Hu MC, Amrhein PC, Nunes EV. Assessment of the relationship between a written measure of empathy and an independently rated interview of Motivational Interviewing. J Subst Abuse Treat. 2018; 87: 56-63.

[40] Cullen MJ, Zhang C, Marcus-Blank B, Braman JP, Tiryaki E, Konia M, Hunt MA, Lee MS, Van Heest A, Englander R, Sackett PR, Andrews JS. Improving Our Ability to Predict Resident Applicant Performance: Validity Evidence for a Situational Judgment Test. Teach Learn Med. 2020; 32: 508-21.

[41] Hartzler B. Adapting the helpful responses questionnaire to assess communication skills involved in delivering contingency management: preliminary psychometrics. J Subst Abuse Treat. 2015; 55:52-7.

[42] Humphris GM, Kaney S. The Objective Structured Video Exam for assessment of communication skills. Med Educ. 2000; 34: 939-45.

[43] Humphris GM. Communication skills knowledge, understanding and OSCE performance in medical trainees: a multivariate prospective study using structural equation modelling. Med Educ. 2002; 36: 842-52.

[44] Kiessling C, Bauer J, Gartmeier M, Iblher P, Karsten G, Kiesewetter J, Moeller GE, Wiesbeck A, Zupanic M, Fischer MR. Development and validation of a computer-based situational judgement test to assess medical students' communication skills in the field of shared decision making. Patient Educ Couns. 2016; 99: 1858-64.

[45] Ludwig S, Behling L, Schmidt U, Fischbeck S. Development and testing of a summative video-based e-examination in relation to an OSCE for measuring communication-related factual and procedural knowledge of medical students. GMS J Med Educ. 2021; 38: Doc70.

[46] Van Dalen J, Kerkhofs E, Verwijnen GM, van Knippenberg-van den Berg BW, van den Hout HA, Scherpbier AJ, van der Vleuten CP. Predicting communication skills with a paper-and-pencil test. Med Educ. 2002; 36: 148-53.

[47] Lievens F, Sackett PR. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. J Appl Psychol. 2012; 97: 460-8.

[48] Karnieli-Miller O, Michael K, Gothelf AB, Palombo M, Meitar D. The associations between reflective ability and communication skills among medical students. Patient Educ Couns. 2021; 104: 92-98.

[49] Lukman H, Beevi Z, Yeap R. Training future doctors to be patient-centred: efficacy of a communication skills training (CST) programme in a Malaysian medical institution. Med J Malaysia. 2009; 64: 51-5.

[50] Cohen: A Power Primer. Psychological Bulletin, 1992, 112:155-159.

[51] Schuwirth LTW, Verheggen MM, van der Vleuten CPM, Boshuizen HPA, Dinant GJ.. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ 2001; 35:348-356.

[52] Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ 2011: 45: 1181–9.

[53] Van Nuland M, van den Noortgate W, van der Vleuten C, Goedhuys J. Optimizing the utility of communication OSCEs: Omit station-specific checklists and provide students with narrative feedback. Pat Educ Couns 2012; 88: 106-112.

[54] Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. Med Educ 1996; 22: 325–34.

[55 ~~54~~] Gillis AE, Morris MC, Ridgway PF. Communication skills assessment in the final postgraduate years to established practice: a systematic review. Postgrad Med J 2015; 91:13-21.

[56 ~~55~~] Figueiredo MN, Krippeit L, Ihorst G, et al. ComOn-Coaching: The effect of a varied number of coaching sessions on transfer into clinical practice following communication skills training in oncology: Results of a randomized controlled trial. PLoS One 2018;13:e0205315.

[57 ~~55~~] Schell JO, Cohen RA, Green JA, Rubio D, Childers JW, Claxton R, et al. NephroTalk: Evaluation of a Palliative Care Communication Curriculum for Nephrology Fellows. J Pain Symptom Manag. 2018;56(5):767–73.e2.

[58 ~~56~~] Bylund CL, Brown R, Gueguen JA, Diamond C, Bianculli J, Kissane DW. The implementation and assessment of a comprehensive communication skills training curriculum for oncologists. Psycho-Oncology. 2010;19(6):583–93.

[59 ~~57~~] Alexander SC, Keitz SA, Sloane R, Tulsky JA. A controlled trial of a short course to improve residents' communication with patients at the end of life. Acad Med. 2006;81(11):1008–12

[60 ~~58~~] Weiland, A, Blankenstein AH, Van Saase JLCM, et al., Training Medical Specialists to Communicate Better with Patients with Medically Unexplained Physical Symptoms (MUPS). A Randomized, Controlled Trial. PloS one. 2015;10(9):e0138342.

[61 ~~59~~] Kramer AWM, Düsman H, Tan LHC, Jansen JJM, Grol RPTM, van der Vleuten CPM. Acquisition of communication skills in postgraduate training for General Practice. Med Educ 2004; 38:158-67.

[62 ~~60~~] Berendonk C, Schirlo C, Balestra G, Bonvin R, Feller S, Huber P, et al. The new final Clinical Skills examination in human medicine in Switzerland: Essential steps of exam development, implementation and evaluation, and central insights from the perspective of the national Working Group. GMS Z Med Ausbild. 2015;32(4):Doc40.

[63 ~~61~~] Brailovwski C, Charlin B, Beausoleil S, Coté S, van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. Med Educ 2001;35:430-436.

[64 ~~62~~] Remmen R, Scherpbier A, Denekens J, Derese A. Hermann I, Hoogenboom R, et al., Correlation of a written test of skills and a performance based test: a study in two traditional medical schools, Med Teacher 2001; 23:29-32.

[65 ~~63~~] Ram P, van der Vleuten C, Rethans JJ, Schouten B, Hobma S, Grol R.. Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice. Med Educ 2002; 33:197-203.

[66 ~~64~~] Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurement 1982; 6:1-11.

[67 ~~64~~] Lievens F, Sackett PR. Video-based versus written situational judgment tests: A comparison in terms of predictive validity. J Applied Psychology 2006; 91: 1181–1188.

[68 ~~65~~] Turner JL, Dankoski ME. Objective Structured Clinical Exams: A Critical Review. Fam Med 2008; 40:574-78.

[69 ~~66~~] Perkins DN, Salomon G. Are Cognitive Skills Context-Bound? Educational Researcher 1989; 18: 16-35.

[70 ~~67~~] Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004; 38: 1006-12.

[71 ~~68~~] Parkes JT, Suen HK, Zimmaro DM, Zappe SM. Structural Knowledge as a Pre-requisite to Valid Performance Assessment Scores. Paper presented at the Annual Meeting of the National Council on Measurement in Education: Montreal, Canada. April, 1999. https://citeseerx.ist.psu.edu/viewdoc/download (accessed 2022, July 8th).

[72 ~~69~~] Grandgenett N, Thompson A. Effects of guided programming instruction on the transfer of analogical reasoning. J Educational Computing Research 1991; 7:293-308.

[73 ~~70~~] Hattie JAC, Donoghue GM. 2016 Learning strategies: a synthesis and conceptual model. Science of Learning 2016; 1:16013.

[74 ~~71~~] Baig LA, Violato C, Crutcher RA. Assessing clinical communication skills in physicians: are the skills context specific or generalizable. BMC Medical Education 2009, 9:22.

[75 ~~72~~] Essers G, van Dulmen S, van Es J, van Weel C, van der Vleuten C, Kramer A. Context factors in consultations of general practitioner trainees and their impact on communication assessment in the authentic setting. Pat Educ Couns 2013; 93: 567-72.

[76 ~~73~~] Peyton JWR. The learning cycle. In: Peyton JWR, editor. Teaching and learning in medical practice. Rickmansworth, UK: Manticore Europe Limited, 1998, p. 13-19.

[77 ~~74~~] Van der Vleuten CPM, Schuwirth LWT, Scheele F, Driesen EW, Hodges B. The assessment of professional competence: building blocks for theory development Best Pract Res Clin Obstet Gynaecol 2010; 24:703–19.

[78 ~~75~~] Cullen MJ, Zhang C, Sackett PR, Thakker K, Young JQ. Can a Situational Judgment Test Identify Trainees at Risk of Professionalism Issues? A Multi-Institutional, Prospective Cohort Study Acad med 2022 May 24th; Online ahead of print.

[79 ~~76~~] Patterson F, Ashworth V, Zibarras L, Coan P, Kerrin M, O'Neill P. Evaluations of situational judgement tests to assess non-academic attributes in selection. Med Educ 2012: 46: 850–68.

Table 1: Description and quality assessment of studies investigating the association between written assessment and performance-based assessment

First author Year Country	Number of participants Background Level of training	Quality assessment of studies	1) Name (or type) of the test 2) Communication topics	Written test characteristics 1) Number of items 2) Stimulus 3) Prompt: type of knowledge 4) Response format 5) Reliability (between items or between raters)	Performance Measu characteristics 1) Setting 2) Scoring/Marking Instrument(s) 3) Reliability (betwe stations/Items or ra
Smith J [39] 2018 USA	97 Counselors from substance abuse community treatment programs Continuous education	Methodology 4 Information 4 Strengths of findings 4	1) Helpful Response Questionnaire (HRQ) 2) Specific CS: motivational interviewing	1) 6 items 2) Context rich - written vignette 3) Behavior-oriented 4) Constructed response 5) ICC =.79	1) 1 randomly select audiotaped segment interaction with a clie **observation time)** 2) MITI 2.0: 2: global (spirit; empathy), 7 c behaviors 3) ICC (on item level .74; rater agreement score: 37%
Cullen M [40] 2020 USA	312 in 21 residency programs Medicine Postgraduate	Methodology 3 Information 4 Strengths of findings 3	1) Situational Judgement Test (SJT) 2) General/basic CS (within professionalism dimensions)	1) 15 scenarios with 7 response options each 2) Context rich - written vignette 3) Knowledge-oriented 4) Selected response 5) α =.86	1) Workplace-based 2) ACGME milestone performance (6 core scores); multisource professionalism asse (MPA; 7 dimensions) retrospective USMLE Step 2 CK, and Step 3) α =.92 (MPA comp
Hartzler B [41] 2015 USA	19 staff clinicians Postgraduate	Methodology 3 Information 4 Strengths of findings 3	1) Helpful Responses Questionnaire adapted for Contingency Management (HRQ-CM) 2) Specific CS: motivational interviewing	1) 6 items 2) Context rich - written vignette 3) Behavior-oriented 4) Constructed response 5) α =.74; ICC (total score) =.96; ICCs (on item level) betw. 0.77 & 0.93	1) 1 audiotaped SP e **(20 min. observatio** 2) Contingency Mana Competence Scale (skill domains) 3) ICCs (on item leve .77 & .89
Humphris G 2000 [42] & 2002 [43] UK	200 (in 2000); 383 (in 2002) Medicine Undergraduate	Methodology 4 Information 4 Strengths of findings 4	1) Objective Structured Video Examination (OSVE) 2) General/basic CS	1) 1 video-based scenario, 3 open answer questions 2) Context rich - video 3) Both knowledge and behavior oriented 4) Constructed response 5) ICC =.94 (Humphris 2000)	1) 4-station OSCE w **resp. 30 minutes ob** **time)** 2) Expert examiners Communication Skill Assessment Scale (I SPs: Global Simulate

					Rating Scale (GSPR 3) not reported
Kiessling C [44] 2016 Germany	72 Medicine Undergraduate	Methodology 4 Information 4 Strengths of findings 4	1) Computer-Based Test including a Situational Judgement Test (CBT) 2) General/basic CS & specific CS: shared decision making	1) Part A 7 items; part B 8 vignettes with 15 items 2) Context poor & context rich - written vignette 3) Knowledge-oriented 4) Selected response 5) α =.58 (total score); part A α =.68; part B α =.41	1) 2 videotaped SP e 2) Berlin Global Rati Observing Patient In Instrument (OPTION 3) α =.83 (BGR: 2 ra cases, 4 items); α =. 2 raters, 2 cases, 10
Ludwig S [45] 2021 Germany	192 Medicine Undergraduate	Methodology 3 Information 4 Strengths of findings 3	1) Video-based Single-choice e-Examination (VSE); SJT 2) General/basic CS & Specific CS: counselling, shared decision making, breaking bad news	1) videos with 30 items 2) Context rich - video 3) Both knowledge and behavior oriented 4) Selected response 5) α =.55 (term 2018) α =.62 (term 2018/19)	1) 5-station OSCE w 2) Self-developed ma checklists (content a communication skills 3) α =.55
Van Dalen J [46] 2002 The Netherlands	133 Medicine Undergraduate	Methodology 3 Information 3 Strengths of findings 3	1) Knowledge test about Communication Skills (KCS) 2) General/basic CS	1) 78 items 2) Context rich - written vignette 3) Knowledge-oriented 4) Selected response 5) α =.62 (Maastricht α =.41; Leiden α =.72)	1) 4 stations with SP **min. observation ti** 2) MAAS-Global 3) α =.49 (Maastricht Leiden α =.24)
Lievens P [47] 2012 Belgium	723 Medicine Undergraduate	Methodology 3 Information 4 Strengths of findings 3	1) Situational Judgement Test (SJT) 2) General/basic CS	1) 30 items 2) Context rich - video 3) Knowledge-oriented 4) Selected response 5) α =.66 (reference to another publication [69])	1) Workplace-based (Overall internship & 2) Internship perform composite; job perfo rating score sheets (participants); Grade Average (GPA) durin studies 3) not reported
Karnieli-Miller O [48] 2021 Israel	66 Medicine Undergraduate	Methodology 4 Information 4 Strengths of findings 4	1) Reflection Evaluation for Learners' Enhanced Competencies (REFLECT) with 4 additional parameters 2) Specific CS: breaking bad news	1) 1 reflective writing 2) Context rich - observation of a real-life encounter 3) Behavior-oriented/reflection 4) Constructed response 5) Not reported	1) 1 videotaped enco SP 2) Breaking Bad New Assessment Schedu SPIKES Questionna 3) α =.91 (BAS); α =.

Lukman H [49] 2009 Malaysia	189 Medicine Undergraduate	Methodology 3 Information 3 Strengths of findings 3	1) Communication Skills Video Assessment (CSVA) 2) General/basic CS	1) 1 Video, 11 short answer questions 2) Context rich - video 3) Both knowledge- & behavior-oriented 4) Constructed response 5) Not reported	1) 1 station History T OSCE **(5 min. obse** 2) Self-developed co skills checklist 3) Not reported

Legend: CS = communication skills; ICC = intraclass correlation; SP = Simulated patients; SJT = Situational Judgement Test; USMLE = United States Medical Licensing Examination; ACGME = Accreditation Council for Graduate Medical Education; MPA = multisource professionalism assessment; BGR = Berlin Global Rating; OPTON = Observing Patient Involvement Instrument; VSE = Video-based Single-choice e-Examination; MSE = Multiple Station Examination; MAAS-Global = Maastricht History-taking and Advice Scoring list; GPA = Grade Point Average; BAS = Breaking Bad News Assessment Schedule; SPIKES = setting - perception – invitation – knowledge – emotion - summary