Comparison of two instruments for assessing communication skills in a general practice objective structured clinical examination

Marc Van Nuland,1 Wim Van Den Noortgate,2 Jan Degryse1 & Jo Goedhuys1

OBJECTIVE In recent decades, there has been increased interest in tools for assessing and improving the communication skills of general practice trainees. Recently, experts in the field rated the older Maas Global (MG) and the newer Common Ground (CG) instruments among the better communication skills assessment tools. This report seeks to establish their cross-validity.

METHODS Eighty trainees were observed by 2 raters for each instrument in 2 standardised patient stations from the final year objective structured clinical examination for Belgian trainee general practitioners. Each instrument was assigned 6 raters.

RESULTS Trainees showed the lowest mean scores for evaluating the consultation (MG7), summarising (MG11), addressing emotions (MG9) and addressing feelings (CG5). Inter-rater $\kappa$ statistics revealed fair-to-moderate agreement for the MG and slight-to-fair agreement for the CG. Cronbach’s $\alpha$ was 0.78 for the MG and 0.89 for the CG. A generalisability study was only feasible for the MG: it was more helpful to increase the number of cases than the number of raters. Agreement between the instruments was examined using $\kappa$ statistics, Bland–Altman plots and multi-level analysis. Ranking the trainees for each instrument revealed similar results for the least competent trainees. Variances between and within trainees differed between instruments, whereas case specificity was comparable. Multi-level analysis also revealed a rater–item interaction effect.

CONCLUSIONS The 2 instruments have convergent validity, but the drawbacks of the CG, which has fewer items to be scored, include lower inter-rater reliability and score variance within trainees.

KEYWORDS comparative study [publication type]; clinical competence/ *standards; *communication; *physician–patient relations; family practice/ *standards; vocational education; Belgium.

Medical Education 2007
doi:10.1111/j.1365-2923.2007.02788.x

INTRODUCTION

The importance of competent doctor–patient communication as central to effective general practice is now widely accepted. It has also been emphasised in the European Definition of General Practice/Family Medicine. Accordingly, more attention has been given in recent decades to evaluating communication skills in the summative assessment of trainee general practitioners (GPs). This is also true in Flanders, the Dutch-speaking part of Belgium. Part of the Flemish objective structured clinical examination (OSCE) is focused on communication skills. Unfortunately, the scores on communication skills within this OSCE are disappointingly low. Interventions to improve communication skills during GP vocational training have had variable results, with some interventions apparently resulting in certain ameliorations, whereas others show no effect in improving certain aspects of GP trainee performance. The quality of the assessment instruments used is crucial to this type of research. The absence of a ‘gold standard’ in medical education, which also applies to the assessment of communication skills, and the issue of context specificity makes it...
Overview

What is already known on this subject

The Maas Global (MG) and the Common Ground (CG) instruments are rated among the better communication assessment instruments.

What this study adds

The MG and CG have convergent validity. The CG measures a single major aspect of communication skills, whereas the MG takes a broader view.

The drawback of the CG, which has a lower number of items, is its lower inter-rater reliability and variance within trainees.

Multi-level analysis reveals an interaction effect between rater and item.

Suggestions for future research

The instruments could be compared in a higher number of stations.

Item-rates interaction could be researched qualitatively.

difficult to assess communication skills. Recent reviews of communication skills assessment instruments show that numerous instruments exist, but they differ in many ways.9–11 Instruments vary in the number of items to be scored, ranging from few (e.g. the Common Ground instrument12) to many (e.g. the LIV-MAAS scale13). Recently, a consensus panel of 6 experts in the field evaluated 15 instruments for assessing doctor–patient communication.9 They rated the Maas Global (MG)14,15 and the Common Ground (CG)12 instruments among the better tools with regard to their overall value, usability and psychometric properties. Both instruments are based on global ratings and earlier research has suggested that they are more reliable than detailed and lengthy checklists.11 Experience with the MG goes back about 20 years16 and is supported by many validation studies.14,15,17 The CG, by contrast, was developed only recently.12 It seems promising in that it focuses more on what are believed to be essential communication skills18 and in the lower number of items to be scored. This research report aims to compare the CG with the better known MG in terms of item scores, internal consistency, generalisability and agreement between raters and between instruments. By investigating the properties of both instruments, this report may produce some evidence to guide the choice between them.

METHODS

We selected 2 stations with a focus on communication from the final year OSCE for Flemish general practice vocational trainees and assessed all participating trainees. In the first station (P), the standardised patient (SP) impersonated a 36-year-old with an unplanned pregnancy. The second station (O) presented a 50-year-old obese man with ambivalence towards efforts to lose weight. Each of these 8-minute stations was directly observed by 4 raters: 2 using the Dutch MG19 and 2 using the translated CG. The MG contains 13 global ratings for the assessment of communication skills, scored on a 7-point Likert scale of 0–6. For both stations, only 11 of the 13 items from the MG instrument were applicable; the scores for Follow-up Consultation and Physical Examination were irrelevant in the context of these stations. The CG consists of 7 global ratings scored on a 5-point Likert scale of 1–5. In order to compare both instruments in the way they evaluate trainee competence, we chose the middle of both scales (3) as the cut-off for distinguishing between insufficient and sufficient competence. Twelve raters were recruited by advertisement among final-year psychology students and recently graduated psychologists. All raters were given a 3.5-hour training session on 1 of the assessment instruments, 6 on the MG and 6 on the CG. The SPs had participated in previous OSCEs and were trained by faculty members for 3 hours in performing the selected stations (O or P). Both raters and SPs were paid.

We used sas software for most of the statistical analyses (descriptive statistics, factor analysis, Cronbach’s α, inter-rater κ calculation, multi-level analysis), genova for generalisability analysis and MedCalc to construct Bland–Altman plots. We tested null hypotheses using 2-sided significance tests with a 5% level of significance.

© Blackwell Publishing Ltd 2007. MEDICAL EDUCATION 2007
RESULTS

In total 80 vocational trainees were observed, yielding 4 × 80 total scores per instrument. Raters differed in the number of trainees they observed, which ranged from 20 to 80, with a mean of about 53 trainees. Missing item scores were limited to 63 of 5760 (1.1%) for all trainees.

Comparing the instruments for item content and observed scores

Each instrument has its own set of items, shown in the first column of Table 1. The remaining columns summarise the distribution of item scores. On average, our cohort of trainees showed insufficient competency (item score <3) when using the MG in Evaluating the Consultation (MG7), Summarising (MG11) and Dealing with Emotions (MG9). When using the CG, trainees showed insufficiency only in Addressing Feelings (CG5).

Internal consistency

In order to compare the instruments regarding internal consistency, we first calculated Cronbach’s $\alpha$, which equalled 0.78 for the standardised items of the MG and 0.89 for the CG items. This indicates acceptable internal consistency for both. As the last item in the CG (CG7) relates to Global Interview Performance, we recalculated Cronbach’s $\alpha$ without CG7, lowering its value from 0.89 to 0.85.

Secondly, internal consistency for each instrument was investigated by exploratory factor analyses. The extent to which major factors explain common variance within an instrument reflects its internal consistency. Factor analysis of the MG revealed 1 major factor (F1: Patient-oriented Problem Analysis and Explanation) with an eigenvalue of 3.09, which explained 60% of the common variance within the MG scores. By contrast, the 1 major factor we found for the CG (F2: Patient-centred Communication), with an eigenvalue of 3.35, explained a greater proportion (86%) of its common variance. Factor loadings for both major factors are shown in Table 2. Other factors in the analysis of both instruments were not retained as the first factor already explained a substantial portion of the common variance, and as the eigenvalues of the second and subsequent factors were considerably smaller than the eigenvalue of the first factor. This was clearly confirmed in a Cattell’s Scree Plot.20 Moreover, when more factors were retained, they appeared to be difficult to interpret, even after rotation.

<table>
<thead>
<tr>
<th>Item content</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maas Global</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1: Communication skills for each separate phase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MG1 Opening</td>
<td>4.13</td>
<td>0.74</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>MG3 Clarification of the reason for encounter</td>
<td>3.02</td>
<td>1.49</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>MG5 Evaluation and diagnosis</td>
<td>3.98</td>
<td>0.73</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>MG6 Management and arrangements</td>
<td>4.13</td>
<td>0.97</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>MG7 Evaluation of consultation</td>
<td>1.67</td>
<td>1.78</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Section 2: General communication skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MG8 Exploring</td>
<td>4.05</td>
<td>0.90</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>MG9 Dealing with emotions</td>
<td>2.89</td>
<td>1.52</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>MG10 Providing information</td>
<td>4.15</td>
<td>0.88</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>MG11 Summarising</td>
<td>2.66</td>
<td>1.77</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>MG12 Structuring</td>
<td>3.77</td>
<td>0.86</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>MG13 Empathy and flexibility</td>
<td>4.02</td>
<td>1.05</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td><strong>Common Ground</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG1 Rapport building</td>
<td>3.34</td>
<td>0.75</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>CG2 Information management</td>
<td>3.11</td>
<td>0.78</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>CG3 Eliciting all agenda items</td>
<td>3.01</td>
<td>0.80</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>CG4 Active listening for full understanding of ideas, concerns, and expectations</td>
<td>3.32</td>
<td>0.80</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>CG5 Addressing feelings with patients</td>
<td>2.87</td>
<td>0.92</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>CG6 Reaching common ground</td>
<td>3.21</td>
<td>0.82</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>CG7 Global interview performance</td>
<td>3.22</td>
<td>0.82</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

* Calculations of means and standard deviations are based on 4 observations, 2 in each station
SD = standard deviation
Generalisability

Generalisability (G) analysis\textsuperscript{21} showed acceptable reliability of the MG scores. Raising the number of stations increased reliability more than raising the number of raters per station. As a result of the low variance of some scores, it was technically impossible to complete a G analysis of the CG.

Agreement between raters

The mean inter-rater correlation was \( r = 0.49 \) for the MG item scores and \( r = 0.26 \) for the CG item scores. The mean weighted inter-rater \( \kappa \) for the MG was 0.56 (range 0.31–0.74) in the O station and 0.35 (range 0.14–0.56) in the P station. Equivalent mean \( \kappa \) statistics for the CG were, respectively, 0.23 (0.04–0.51) and 0.19 (0.03–0.46).

Agreement between the 2 instruments

To investigate how the 2 instruments agree, we looked at Bland–Altman plots\textsuperscript{22,23} and at their correlation matrix, calculated \( \kappa \) statistics and made empirical Bayes estimates based on the results of multi-level analyses.

1 In order to construct Bland–Altman plots, we transformed the scores into percentages (Fig. 1). These plots give insight about whether agreement depends upon the magnitude of the measurement and show that for weak trainees (an average of < 3, or <50\%), scores tended to be higher with the MG than with the CG, whereas for stronger trainees, scores tended to be lower with the MG than with the CG.

2 The correlation matrix of the items of both instruments does not reveal any relevant correlation between MG items and CG items (all \( r < 0.32 \)). The correlation between the factors computed in the factor analysis of the MG and CG measured \( r = 0.36 \) (\( P < 0.0001 \)), suggesting that they assess rather different dimensions of communication.

3 As we had defined scores of <3 as insufficient ability (fail) and scores \( \geq 3 \) as sufficient ability (pass), the \( \kappa \) between instruments on the pass or
fail decision was 0.37 for the O station and 0.40 for the P station, indicating slight-to-fair agreement.  

Multi-level analysis (MLA) was used to further explore the relationship between the 2 instruments. In general, multi-level models are used to analyse clustered data. Characteristics of the clusters, as well as of the units within the clusters, can be used as independent variables in describing the variation in 1 or more dependent variables. In our case, we have item scores that are clustered in persons. We used a multivariate model, which treated the item scores of the CG and the MG as 2 dependent variables. Multi-level analysis showed that the 2 instruments had different mean scores (MG 3.50, CG 3.16), as well as a different variance within trainees (MG 1.99, CG 0.61) and between trainees (MG 0.10, CG 0.072). The MLA also revealed an interaction effect between trainees and items: trainees may score relatively strongly or weakly on particular items compared with other items.

Multi-level analysis confirmed case-specificity as MG scores on the obesity station were on average 0.15 lower than on the pregnancy station ($P = 0.0015$), whereas CG scores were only 0.06 lower ($P = 0.07$). Regarding this interaction effect between instrument and station, the difference between the MG and CG was not statistically significant ($P = 0.12$).

Multi-level analysis further revealed significant differences in the way raters gave scores for each instrument ($P < 0.0001$ for the MG and CG). It also revealed an interaction effect between rater and item ($P < 0.0001$), meaning that individual raters may tend to give higher or lower scores on particular items in comparison with other raters.

As in practice scores will usually be based on 1 instrument only, we also performed 2 univariate MLAs (1 for each instrument) in order to calculate the empirical Bayes estimates of trainees’ skills. A correlation of $r = 0.59$ ($P < 0.0001$) was found between the estimates of MG and CG scores (Fig. 2).

It was possible to rank the trainees based on the estimates of their total scores for the 2 stations. A Spearman’s rank correlation coefficient of $r = 0.55$ ($P < 0.0001$) was found. A closer look at the data for this analysis reveals that of the 8 trainees in the first decile of the MG, 5 were in the first decile and 7 were in the first 2 deciles of the CG. Similarly, of those trainees in the first decile of the CG, 5 were in the first decile and 7 in the first 2 deciles of the MG.

Conversely, only 3 of the trainees from the last decile of the MG were found in the last 2 deciles of the CG and 3 from the last decile of the CG were found in the 2 last deciles of the MG. These findings suggest that the 3 instruments agree more in selecting less competent trainees than in selecting the best.

**DISCUSSION**

In this study, we cross-validated 2 communication assessment instruments in a standardised setting (OSCE). Bland–Altman plots and interinstrument $\kappa$ calculation revealed a slight-to-fair agreement between instruments when used in 1 station. Multi-level analysis showed a correlation of 0.59 between the estimates of individual trainees’ total scores on both instruments, based on the observations in 2 stations. These are arguments for convergent validity of the 2 instruments. To our knowledge, this study is the first to compare 2 communication assessment instruments in a simultaneous direct observation setting. To us, this experimental method of comparing assessment instruments seems to be more valid than a comparison based on ratings by experts, such as in the study by Schirmer et al. Three aspects will be considered in the further discussion of our study: study design; study results, and suggestions for further research.

**Study design**

Choosing a standardised setting for our study allowed us the advantage of avoiding the bias caused by the patient variance that would occur in real patient consultations, but the disadvantage of measuring...
competence rather than performance. Considering the purpose of our study, however, this disadvantage is of minor importance. Furthermore, by applying both instruments simultaneously in direct observation and by deploying 2 raters per instrument, we maximised standardisation and minimised bias resulting from patient variance. Bias resulting from missing data seems rather improbable as only 1.1% of item scores were missing and a closer look at the missing scores did not reveal a link to a particular item, rater or trainee. Restricting our study to only 2 stations had the disadvantage of providing less information concerning generalisability and as such, the respective conclusions must be regarded with caution. Restrictions in budget, time and manpower made it impossible to expand our experiment over more than 2 stations.

We chose to engage psychologists to rate communication skills for 2 reasons. Firstly, as they are not doctors, their scores on communication will be less influenced by ratings on the medical content of the consultation, a potential source of bias. Secondly, psychologists are familiar with using rating scales when observing behaviour and human interaction. As such, the 3.5-hour experiential training sessions ought to have been sufficient. A disadvantage of engaging psychologist raters is that we remain uncertain whether doctors, the usual raters of OSCE stations, would have produced equal results and conclusions.

**Study results**

When comparing the MG and CG instruments, the different variances cannot be attributed solely to the different scales. If the 5-point scale from the CG were to be extended to a 7-point scale, the standard deviation (SD) would rise by a factor of $7/5$ and variance by a factor of $1.95 \times \left(\frac{7}{5}\right)^2$. The variance within trainees for the CG ($= 0.61$) would rise to 1.20, which is lower than the MG variance of 1.99. However, the variance between trainees for the CG ($= 0.072$) would rise to 0.14, which is higher than the MG variance of 0.10. This lower variance between trainees in the MG may contribute to the decreasing slope of the Bland–Altman plots.

Ranking trainees according to their scores on 2 stations revealed highly similar results for the weakest trainees on both instruments, suggesting that the MG and CG are comparable in selecting trainees in need of additional support regarding their communication skills. Nevertheless, further research is needed because this study included only 2 OSCE stations and, as a result of case specificity, this number is too low to infer generalisability. Cushing’s overview mentions that 8–24 stations are necessary in order to obtain generalisability, whereas Lang et al. concluded in their study of the CG that only 5 cases were needed to achieve a generalisability coefficient $\geq 0.80$. Our study could not confirm the results of the Lang study, yet the generalisability we found for the MG in our study was more consistent with findings in previous research.

The mean item scores we found on the MG for GP vocational trainees in their final year are very similar to those in the study by van Dalen et al. which are on average higher than those found by Kramer et al. (Fig. 3). It is striking that in both Dutch studies, as in ours, final year GP trainees had the weakest score on the item Evaluating the Consultation. This item also seems to have a low factor loading on the major

![Figure 3](image-url)
factor. This suggests that this item measures an aspect of low relevance to the rest of the scale.

Interobserver agreement for item scores measured by the weighted \( \kappa \) ranged from 0.35 (P) to 0.56 (O) for the MG, coming into the category of fair-to-moderate agreement.\(^{24} \) For the CG these were 0.19 (P) and 0.23 (O), corresponding to slight (P) and fair (O) agreement. This rather low agreement was also found in the study by van Thiel \textit{et al.}\(^{14} \) which used a former version of the MG. The inter-rater agreement we found for the CG was lower than that reported by Lang \textit{et al.}\(^{12} \) who used only 2 strictly selected raters observing videos of 100 students carrying out 4 SP interviews. We, by contrast, used 6 raters per instrument and no pair of raters observed more than 20 of the same trainees.

An interesting result of the MLA was that it showed an interaction effect between raters and items: some raters tended to give relatively high or low scores on specific items, compared with others. This aspect may not be apparent in studies that rely on a small number of raters.

The 2 instruments seemed to differ in terms of internal consistency. The CG7 item (Global Interview Performance) could bias the Cronbach’s \( \alpha \) calculation as it is not a ‘true’ item. As such, calculating \( \alpha \) without CG7 may give a result that more accurately represents internal consistency. Nevertheless, Cronbach’s \( \alpha \) remained higher for the CG than for the MG, although the number of items was smaller.

This result, as well as the higher proportion of variance explained by the first factor from our factor analysis, suggests that the CG measures 1 major aspect of communication skills in particular.

Further research

This study proved to be valuable in gaining better understanding of the MG and CG and showed the extent of agreement between them. Analogous studies with more and different stations (e.g. focusing on specific GP communication tasks, such as breaking bad news, counselling, handling specific conditions) are needed to confirm our results. Our study suggests that increasing the number of stations enhances generalisability better than increasing the number of raters per station. Qualitative research focused on raters may contribute to a better understanding of item–rater interaction.

CONCLUSIONS

This study cross-validated the MG and CG instruments for use in 2 OSCE stations. It revealed that they are comparable regarding case specificity and in tracing weak trainees.

By contrast, the instruments proved to be different in terms of item content, internal consistency (CG > MG), variance of scores within trainees (MG > CG), variance between trainees (MG < CG) and inter-rater reliability (MG > CG). These results may aid in choosing between the 2 instruments, depending on the aim or context of the assessment (e.g. formative or summative; global communication skills or specific skills such as addressing feelings, structuring the consultation or finding common ground).

Multi-level analysis showed a significant item–rater interaction effect for both instruments, shedding new light on the complexity of inter-rater agreement in communication assessment.

Contributors: MVN developed the study’s protocol, trained the raters, organised the data collection, carried out the data analysis, and prepared and edited the manuscript. WVDN carried out data analysis and critically revised the manuscript. JDG raised the original idea of the study design, carried out the generalisability study and revised the manuscript. JG contributed to the study protocol, trained the raters, carried out the data analysis and critically revised the manuscript.

Acknowledgements: we would like to thank the staff of the Centre for Postgraduate General Practice Vocational Training (Interuniversitair Centrum voor HuisartsenOpleiding, ICHO) and the trainees who took part in the objective structured clinical examination.

Funding: none.

Conflicts of interest: none.

Ethical approval: not required.

REFERENCES


Received 11 June 2006; editorial comments to authors 15 October 2006; accepted for publication 20 February 2007