Reliability of the assessment of consultation skills of general practice trainees with the MAAS-Global instrument

Marcel E. Reinders
Annette H. Blankenstein
Harm W.J. van Marwijk
Dirk L. Knol
Paul Ram
Henriëtte E. van der Horst
Henrica C.W. de Vet

submitted
Abstract

Objectives
The training and assessment of consultation skills is high on the agenda of vocational training institutes for general practice. There is a need for valid and reliable instruments to assess consultation skills in ‘real-life’ settings. We investigated the number of assessors/observations needed for reliable assessments of consultation skills of GPTs with the MAAS-Global instrument.

Methods
Eight teachers at the VU medical center in Amsterdam attended a short training course on the use of the MAAS–Global instrument and assessed the consultation skills of 53 GPTs during 176 randomly allocated video-taped consultations (102 with standardized patients and 74 with real patients). All consultations were assessed twice by two independent teachers. Reliability was estimated with the generalisability coefficient (GC). Mixed model regression analysis was used to demonstrate the effects of ‘individual assessor’ and ‘type of consultation’ on the MAAS-Global scores. The teachers completed a questionnaire to evaluate their perceived competence.

Results
The inter-observer variation was the most important component of variance. An acceptable GC was easier to obtain using real-life patients than using standardized patient consultations: using real patients, for GC>0.8 two assessors and nine consultations were required; using standardized patients, for GC>0.7 three assessors and 30 consultations were needed. The difference in MAAS-Global outcome scores between the individual assessors was statistically significant. Consultations with standardized patients were scored significantly higher than real-patient consultations. The teachers felt sufficiently competent in using the MAAS-Global instrument to assess the consultation skills of the GPTs.
Discussion
Despite the perceived competence of the teachers, to achieve acceptable levels of reliability in the assessment of the consultation skills of GPTs, multiple observations are required, as indicated by the results of this study. We recommend the assessment of real-patient consultations instead of consultations with standardized simulation patients.
Introduction

Adequate communication skills of practicing physicians are related to several important health outcomes, such as emotional health, symptom resolution, function and physiologic measures\(^1\). There is strong evidence that consultation skills are correlated with patient satisfaction, maintenance of relationships, and therapy compliance\(^2\). For this reason, much attention is paid to teaching consultation skills to experienced clinicians\(^3\), as well as in the (general practice) postgraduate training\(^4-7\).

The importance of assessment in the process of learning these skills is widely accepted\(^8\). Assessment of individual progress in the acquisition of consultation skills by general practice trainees (GPTs) during their three years of vocational training in the Netherlands is becoming more systematic, by means of successive observations of video-taped consultations\(^9\).

The MAAS-Global is the most commonly used instrument to assess consultation skills among the vocational institutes for general practice in the Netherlands. This instrument was developed at the Maastricht University as an educational tool to give feedback to medical students and GPTs\(^10\). The reliability of this instrument to assess consultation skills of undergraduate medical students has been found to be good, at least under standardized conditions with experienced assessors\(^11\). Using the generalisability theory, reliability of the assessment with the MAAS-Global instrument was expressed as a function of the number of observed consultations and assessors: two assessors and eight consultations were required for a sufficiently reliable assessment\(^12\).

What is less known, however, is how reliable assessments with the MAAS-Global instrument are under less controlled and more real-life conditions (e.g. teachers newly trained, and GPTs selecting one of their recently video-taped consultations) that reflect daily practice in a vocational training institute. Individual differences in the scoring of the teachers reduce the objectivity and reliability of the assessments. If there are multiple assessors, taking the mean of the different scores will reduce the inter-observer variability. However, in more widespread use, or in a formative setting ('education') rather than in a
summative setting (‘examination’), there will probably not always be sufficient manpower for multiple assessments of the consultations. Furthermore, in contrast to standardized consultations, the video-taped GPT consultations in daily practice will vary in content, technical quality and duration.

This is why we conducted a study to assess the reliability of a test procedure with the MAAS-Global instrument in a setting resembling a real-life GP training institute, i.e. of less standardized conditions. Teachers from the VU medical center attended a short training course in the use of the MAAS-Global instrument. Subsequently, they assessed a mixed set of standardized (simulated) patient consultations and real-patient consultations. Their perceived competence was assessed afterwards.

**Aim of the research**

- To assess the reliability of the test method with the MAAS-Global instrument; and to express the generalisability in terms of the number of observations and assessors needed to achieve reliable assessments.
- To assess the differences in MAAS-Global outcome scores between the teachers and between the types of consultations (standardized versus real-patient consultations).
- To assess the perceived competence of the teachers in the use of the MAAS-Global instrument.

**Methods**

**General design and subjects**

In this observational study, eight teachers from the general practice vocational training staff at the VU medical center in Amsterdam, were trained in the use of the MAAS-Global instrument. This instrument consists of 13 case-independent items, measured on a 7-point scale, referring to communication skills, both in the specific phases of the consultation and in the entire consultation (range 0-6:
0=absent, 6=excellent). Two items were not used: ‘first consultation or follow-up’ and ‘physical examination’. The mean total score was calculated, resulting in a total score ranging from 0 to 6. The standard level of adequate performance of GPs has been set at 3.1 (Angoff procedure and borderline regression)\(^{13}\). The teachers subsequently assessed consultation skills by observing video-taped consultations of GPTs. All consultations were scored twice, randomly, by two independent teachers. In the end, all eight teachers completed an evaluative questionnaire about their perception of their competence in assessing GPT consultation skills with the MAAS-Global instrument.

**Educational training**

The training of the teachers in the assessment of consultation skills with the MAAS-Global instrument consisted of: studying the manual; homework assignments (scoring standardized video-taped consultations); and an instruction meeting (during a half-day) supervised by one of the leading experts in this field (PR). The teachers discussed their findings with the expert, and received feedback on their scoring. At the end of the study period the teachers attended an evaluative meeting (2 hours), in which feedback was given to the teachers and the collective and individual results were discussed.

**Materials and allocation scheme**

The video-taped consultations were derived from a patient feedback project, in which 53 first-year GPTs at the VU medical center in Amsterdam were twice visited in their general practices by standardized (simulated) patients, who made video-recordings of the consultations (n=102)\(^{14}\). The simulated patients (three professional actors) were trained to play six comparable scenarios of moderate complexity: the (simulated) patients not only had a physical complaint, but also had a ‘second layer’ anxiety or agenda (complete scenarios available on request from the first author). They were randomly assigned to visit each GPT. In the same period, the GPTs provided one of their recent video-recordings of a real-patient consultation (n=74). Both the GPTs and the real patients gave their consent for the video recordings to be used for educational and research purposes.
All consultations were assessed twice. For the first assessment, the 176 videotaped consultations were allocated to the eight teachers (22 each), according to a computerised random-number generator. For the re-assessment, all consultations were re-allocated equally to different assessors, according to a balanced incomplete block design\(^{15}\), as is shown in Table 1. Different consultations (cases), with different GPTs, were observed by different assessors ('fully-nested design'). The teachers were blinded for each others’ scores.

### Table 1. Allocation schemes (Balanced Incomplete Block Design)

#### 1a. First 176 assessments

<table>
<thead>
<tr>
<th>consultations</th>
<th>assessors (teachers)</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=(1-3)</td>
<td>1A 2A 3A 4A 5A 6A 7A 8A</td>
<td></td>
</tr>
<tr>
<td>B=(4-6)</td>
<td>1B 2B 3B 4B 5B 6B 7B 8B</td>
<td></td>
</tr>
<tr>
<td>C=(7-9)</td>
<td>1C 2C 3C 4C 5C 6C 7C 8C</td>
<td></td>
</tr>
<tr>
<td>D=(10-12)</td>
<td>1D 2D 3D 4D 5D 6D 7D 8D</td>
<td></td>
</tr>
<tr>
<td>E=(13-15)</td>
<td>1E 2E 3E 4E 5E 6E 7E 8E</td>
<td></td>
</tr>
<tr>
<td>F=(16-18)</td>
<td>1F 2F 3F 4F 5F 6F 7F 8F</td>
<td></td>
</tr>
<tr>
<td>G=(19-22)</td>
<td>1G 2G 3G 4G 5G 6G 7G 8G</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=22 n=22 n=22 n=22 n=22 n=22 n=22 n=176</td>
<td></td>
</tr>
</tbody>
</table>

#### 1b. Second 176 (re-)assessments

<table>
<thead>
<tr>
<th>consultations</th>
<th>assessors (teachers)</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A=(1-3)</td>
<td>- 1A 1B 1C 1D 1E 1F 1G</td>
<td></td>
</tr>
<tr>
<td>B=(4-6)</td>
<td>2G - 2A 2B 2C 2D 2E 2F</td>
<td></td>
</tr>
<tr>
<td>C=(7-9)</td>
<td>3F 3G - 3A 3B 3C 3D 3E</td>
<td></td>
</tr>
<tr>
<td>D=(10-12)</td>
<td>4E 4F 4G - 4A 4B 4C 4D</td>
<td></td>
</tr>
<tr>
<td>E=(13-15)</td>
<td>5D 5E 5F 5G - 5A 5B 5C</td>
<td></td>
</tr>
<tr>
<td>F=(16-18)</td>
<td>6C 6D 6E 6F 6G - 6A 6B</td>
<td></td>
</tr>
<tr>
<td>G=(19-22)</td>
<td>7B 7C 7D 7E 7F 7G - 7A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=22 n=22 n=22 n=22 n=22 n=22 n=22 n=176</td>
<td></td>
</tr>
</tbody>
</table>
Data-analysis
Reliability
The generalisability coefficient (GC) assesses reliability while taking various sources of error (or variance components) into account. This information makes it possible to calculate how many observations and how many assessors are needed to achieve a predefined level of reliability\textsuperscript{16,17}.

We used a mixed-model analysis to calculate the components of variance ($\sigma^2$), by means of ANOVA. The following components of variance were estimated: variance between the quality of consultations (GPT), variance between multiple consultations (video-tape) of GPTs (consult/GPT), and the variance between the assessors assessor (teacher) (assessor/consult/GPT). We calculated the GC with the following generalisability formula (in which $\sigma^2$ = the variance component, and $n$ = the number of assessors):

\[
GC = \frac{\sigma^2_{GPT}}{\sigma^2_{GPT} + \frac{\sigma^2_{\text{consult} / \text{GPT}}}{n_{\text{consult}}} + \frac{\sigma^2_{\text{assessor} / \text{consult} / \text{GPT}}}{n_{\text{consult}} \cdot n_{\text{assessor}}}}
\]

By increasing the number of consultations per GPT or by increasing the number of assessors who judge these consultations, and taking the mean value of these MAAS-Global scores, the variance components are divided by the respective $n$ and the GC becomes higher. High-stakes assessments (summative assessments) generally require a GC of 0.8; otherwise (formative assessments) 0.7 is considered sufficient\textsuperscript{18}. We calculate how many consultations are needed or how many assessors, to obtain an acceptable value for the GC.

Influence of assessor and consultation characteristics
The effect of ‘assessor’ and ‘standardization of consultations’ on the MAAS-Global scores was measured. A mixed-model regression analysis, with ‘assessor’ as random factor, was performed to demonstrate the effect of the individual assessors on the MAAS-Global score (as dependent value), and tested for significance with the likelihood ratio. Univariate linear regression analysis was
applied to test whether the MAAS-Global scores were affected by: the differences in type of consultation (standardized patient versus real patient), baseline characteristics of the teachers (background [GP versus behavioural scientist], gender and work experience of the teacher [in years]).

Evaluation of perceived competence
The evaluation questionnaire contained open format questions concerning the time spent on assessment, and the self perceived competence in the use of the MAAS-Global instrument. The evaluation questionnaire is available on request (MR).

Results

Characteristics of the assessors
The teachers formed a mixed team of general practitioners and behavioural scientists/psychologists (ratio 3/5), gender (male/female: ratio 3/5), and teaching experience (mean 11.3 years (SD 10.4), range 1-32 years). Their mean assessment time per consultation was 29.7 minutes (SD 11.2).

Reliability: aspects of inter-observer variability
When all consultations are taken into account, the calculated scores for the variance components were: GPT (0.08), consultations/GPT (0.36), and assessors/consultations/GPT (0.55). This latter term represents the inter-observer variation.

When the components of variance were estimated for real-patient consultations and standardized consultations separately, it appeared that they were very different. Therefore, in Table 2, the components of variance and the generalisability coefficients were set as a function of the number of consultations and assessors, for real-patient consultations and standardized-patient consultations separately. The variance between the teachers was the greatest source of the total variance, and comparable between the two types of
consultations (0.61 and 0.63, respectively). The GC for the situation when only one consultation is assessed by one assessor (teacher) was 0.22 for the real-patient consultations and 0.05 for the standardized-patient consultations, as can be seen in the upper left corner of Table 2. The GC can be improved by increasing using the mean value of four assessors to 0.41 for the real-patients and to 0.09 for the standardized patients (first row in Table 2). By increasing the number of consultations per GPT (represented in the columns in Table 2) one can see that the value of GC (0.8) is reached by nine consultations per GPT using two assessors, or by 15 consultations when the consults are scored by only one assessor. With three or four assessors, seven consultations are needed. For the standardized patients a GC>0.8 cannot easily be reached: more than four assessors scoring more than 30 consultations per patients are needed. A GC of 0.7 can be reached when three assessors score 30 consultations or four assessors score 25 consultations.

**Effect of ‘assessor’ and ‘standardization of consultations’ on the MAAS-Global scores**

The mean MAAS-Global score for all observations was 3.22 (SD 0.99, range 0.64–5.64), and the mean scores varied among the eight teachers, range 2.59 (SD 0.63)–3.75 (SD 1.41), as is shown in Figure 1.

Mixed-model analysis, with ‘teacher’ as random factor and MAAS-Global score as dependent value, demonstrated the effect of the individual teachers on the MAAS-Global score (Table 3). The likelihood ratio test showed that this random effect was statistically significant ($\chi^2=43.9$ (df=1), $p<.0001$). The background of the teachers (GP or behavioural scientist), gender and work experience as a teacher did not affect the MAAS-Global score (univariate analysis, $p=.97$, $p=.81$ and $p=.11$, respectively), as shown in Table 3.

Standardized-patient consultations were related to higher mean MAAS-Global scores (3.51 [SD 0.91]) than were real-patient consultations (2.83 [SD 0.95], $p<.0001$, $t$-test for independent values).
Evaluation of perceived competence

The teachers considered the educational training to be adequate to learn how to use the MAAS-Global (reported by seven teachers, one already had some previous experience). Perceived competence in the assessment of consultation skills was reported by seven teachers (one still experienced some difficulties), which was experienced immediately after the introduction training (by four) or after the first 22 assessments (by two), or unknown (by two). Three teachers specifically appreciated the feeling that they had more ‘grip’ on their assessments, and had developed a more structured approach.

Table 2. Reliability: generalisability coefficients as a function of the number of observations and assessors

<table>
<thead>
<tr>
<th>No. of consultations</th>
<th>No. of assessors</th>
<th>No. of assessors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Real-patient consultations</td>
<td>Standardized-patient consultations</td>
</tr>
<tr>
<td>σ² (GPT)= 0.22</td>
<td>σ² (GPT)= 0.05</td>
<td></td>
</tr>
<tr>
<td>σ² (consultations)= 0.17</td>
<td>σ² (consultations)= 0.32</td>
<td></td>
</tr>
<tr>
<td>σ² (assessor)= 0.61</td>
<td>σ² (assessor)= 0.63</td>
<td></td>
</tr>
</tbody>
</table>

σ² = the variance component; GPT = general practice trainee. Example calculation of a general coefficient (GC) for 2 assessors and 9 consultations for real-patients in the formula:

\[
GC = \frac{\sigma^2_{GPT}}{\sigma^2_{GPT} + \frac{\sigma^2_{consult/GPT}}{n_{consult}} + \frac{\sigma^2_{assessor/consult/GPT}}{n_{consult} \cdot n_{assessor}}} = \frac{0.22}{0.22 + 0.17/9 + 0.61/(9 \times 2)} = 0.81
\]
Table 3. Results of univariate linear regression analysis to assess the influence of assessor and consultation characteristics on the MAAS-Global scores

<table>
<thead>
<tr>
<th>Teacher (gender, background, working experience in years)</th>
<th>MAAS-Global score Mean (SD)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (F, MSc, 1)</td>
<td>2.96 (0.62)</td>
<td>&lt;.0001¹</td>
</tr>
<tr>
<td>2 (M, MD, 19)</td>
<td>2.89 (0.91)</td>
<td></td>
</tr>
<tr>
<td>3 (F, MD, 15)</td>
<td>3.75 (1.41)</td>
<td></td>
</tr>
<tr>
<td>4 (F, MSc, 2)</td>
<td>3.74 (0.79)</td>
<td></td>
</tr>
<tr>
<td>5 (M, MSc, 32)</td>
<td>3.51 (0.91)</td>
<td></td>
</tr>
<tr>
<td>6 (F, MSc, 7)</td>
<td>2.59 (0.63)</td>
<td></td>
</tr>
<tr>
<td>7 (M, MSc, 8)</td>
<td>3.26 (0.72)</td>
<td></td>
</tr>
<tr>
<td>8 (F, MD, 6)</td>
<td>3.68 (1.06)</td>
<td></td>
</tr>
</tbody>
</table>

Teacher characteristics

| GP (n = 3) | 3.24 (1.19) | .81 |
| MSc (n = 5) | 3.21 (0.84) |      |

| Male (n = 3) | 3.22 (0.88) | .96 |
| Female (n = 5) | 3.23 (1.05) |      |

Work experience (in years) .11

Acquired experience

| Assessments (n = 176) | 3.18 (1.03) | .38 |
| Re-assessments (n = 176) | 3.27 (0.94) |      |

Consultation

| Standardized-patient | 3.51 (0.91) | <.0001 |
| Real-patient         | 2.83 (0.95) |        |

¹Likelihood ratio test; F=female, M=male; MD=Medical Doctor; MSc=Master of Science (psychology, behavioural sciences)
Figure 1. Box and whisker plot of mean MAAS-Global scores per teacher

Boxes represent the 25, 50 (median) and 75 cumulative relative frequencies (percentiles); whiskers represent the range; dots represent outliers
Chapter 8

Discussion

Main findings
The teachers felt competent in using the MAAS-Global instrument to assess the consultation skills of GPTs, even in the early stage of the training, and generally valued its educational potential. However, the inter-observer variation was high: the assessor was a significant factor of variance. Our analysis by means of the generalisability theory demonstrated that the number of consultations (rather than the number of assessors) needed for reliable assessment of the consultation skills of GPTs greatly depended on the type of consultation. Real-patient consultations received significantly lower MAAS-Global scores than standardized-patient consultations, but also showed more variance (0.22 versus 0.05), and therefore less real-patient consultations are needed to be scored for reliable assessments, compared to the standardized-patient consultations. Apparently the quality of the consultation of the GPT differed more among the consultations of the real patients than among the consultations of the standardized patients.

Interpretation
The difference in scoring among the teachers was statistically significant but could not be explained by the different educational background, gender, or work experience of the teachers, because these variables did not seem to correlate with the MAAS-Global scores. However, the number of assessors (8) was too small to allow conclusions. It is remarkable that, when asked, the majority of them were aware of their own tendency to be critical or mild in their appraisals.

The type of consultation (standardized consultations versus real-patient consultations) influenced the height and variation of the outcome score. This might be due to the fact that real-patient consultations are more influenced by a variety in factors such as the length of the consultations, the technical quality of the video-tapes of the consultations, or their medical content (‘sample criteria’). It was surprising to know that real-patient consultations had a higher reliability than standardized patients.
Comparing the results of the performance of the MAAS-Global instrument with other studies is hampered by a difference in standardization of the scoring conditions, for example consultations in daily practice or multiple-station examination\textsuperscript{12,19,20}. It is important to realize that the reliability of the assessments of consultation skills depends on the intrinsic qualities of the MAAS-Global instrument as well as on the test-procedure (‘setting’). It is the latter aspect we focused upon here.

It is likely that the inter-observer variability might be reduced by more intensive training of the assessors. In a previous study it was found that for reliable assessment of the consultation skills of experienced GPs with the MAAS-Global instrument (GC 0.8), two assessors and eight (standardized) consultations were needed\textsuperscript{12}.

**Strength and limitations**

Our study, with teachers from a general practice vocational training staff, reflects the reality of daily practice in the assessment of consultation skills. The teachers experienced the intricate nature of scoring consultations. The assessors were closely monitored, which contributed to the fact that we only had a minimum of missing data.

Our study design included a large number of observations, which were all scored twice, by different assessors, in a randomization scheme that provided a maximum of variability.

Unfortunately, our study design in which video-tapes were assessed and re-assessed did not make it possible to give interim feedback to the teachers on their individual assessment performance (for example by means of benchmarking). It is likely that such interim reinforcement feedback, aimed at self-correction, might have a beneficial effect on reliability.

**Practice implications and future research**

What can we do to optimize the generalisability of assessing the consultation skills of GPTs? Apparently, assessments of real-patient consultations lead to a higher GC than assessments of standardized consultations do.
Demands will be made on the acquired experience of the teachers in similar future projects. It is interesting to examine whether feedback on individual performance of teachers (‘hawks’ and ‘doves’) will confirm or cause adjusted behaviour in the future assessment of consultations. If the individual assessment styles of teachers are already known, in future settings it would probably be wise to make matched pairs of critical and mild assessors (at both ends of the scale), which would then balance out their differences.

**Conclusion**

In conclusion, despite the feelings of competence that the teachers experienced, acquiring acceptable levels of reliability in the assessment of consultation skills remains difficult. The consequences for practical use in a real-life setting in the vocational training institute for GPs are that multiple observations are needed, as indicated by the results of this study.
References


