Effects of item and rater characteristics on checklist recording: what should we look for?

HUBER, Philippe, et al.

Abstract

OBJECTIVE: Examinations based on using standardised patients (SPs) commonly use checklist recordings to evaluate students' clinical performance. This paper examines whether and to what extent item and rater characteristics affect the reliability of history checklist recording in an SP-based assessment. METHODS: Checklist items were reviewed for the presence or absence of 5 item characteristics and a 2-point versus 3-point scoring scale. Agreement between checklist recordings obtained from SPs and clinician-examiners (CEs) were compared by item characteristics, scoring scale and CEs' level of involvement in the assessment. RESULTS: Based on 3179 pairs of recordings, the overall percentage of agreement between SPs and CEs was 83% (kappa = 0.64). Agreement was significantly higher for items scored on a 2-point than on a 3-point scale, and when the CE was also the author and the trainer of the station. After controlling for other factors, item characteristics were only marginally associated with level of interrater agreement. CONCLUSIONS: This study suggests that attention should be paid to specific aspects of checklist [...]
Effects of Item and Raters' Characteristics on Checklist Recording:
What Should One Look For?

MED-2004-0127.R2

Correspondence
Dr Philippe Huber, Unit of Development and Research in Medical Education, Faculty of
Medicine, University of Geneva, 1 rue Michel-Servet, CH-1211 Geneva, Switzerland.
Tel : 00 41 22 379 59 40
Fax : 00 41 22 379 51 22
E-mail : philippe.huber@medecine.unige.ch

Authors
Philippe Huber (PH)
1 Unit of Development and Research in Medical Education, Faculty of Medicine, University
of Geneva, Switzerland.
2 Department of rehabilitation and geriatrics, Faculty of Medicine, University of Geneva,
Switzerland.

Anne Baroffio (AB)
1 Unit of Development and Research in Medical Education, Faculty of Medicine, University
of Geneva, Switzerland.

Eric Chamot (EC)
3 Department of Epidemiology and International Health, University of Alabama, Birmingham

François Herrmann (FH)
2 Department of rehabilitation and geriatrics, Faculty of Medicine, University of Geneva,
Switzerland.

Mathieu R Nendaz (MRN)
1 Unit of Development and Research in Medical Education, Faculty of Medicine, University
of Geneva, Switzerland.
4 Department of Internal Medicine, Faculty of Medicine, University of Geneva, Switzerland

Nu V Vu (NVV)
1 Unit of Development and Research in Medical Education, Faculty of Medicine, University
of Geneva, Switzerland.
Contributions made by all authors

PH was the principal investigator and primary author of the manuscript. He was involved in all aspects of the study. AB was involved in conceiving and designing the study, as well as in acquisition and analysis of data. EC and FH participated in statistical analysis and interpretation of the data. MRN participated in acquisition and analysis of the data. NVV served as senior investigator and participated in all phases of the study. All authors participated in critical revision of the article for important intellectual content and gave it final approval.

Acknowledgments

We would like to acknowledge Dr Arnaud Perrier for his critical review of the manuscript and Prof. Alain Junod for his support for the collection of data.

Funding

None.

Conflict of interest

None.

Ethical approval

None was needed.
Effects of Item and Raters' Characteristics on Checklist Recording:

What Should One Look For?
OVERVIEW BOX

What is already known
Standardized patient-based examinations commonly use checklist recordings to evaluate students’ clinical performance. Several factors affect interrater agreement between standardized patient’s and clinician-examiner’s recordings.

What the study adds
Interrater agreements on history checklist recordings are mainly influenced by the scoring scales used, clinician-examiners' level of involvement in the assessment, and to a lesser degree characteristics of the item construct.

Suggestions for further research
Different features in checklist development and checklist recording training need to be taken into consideration in order to ensure its validity and reliability when a standardized patient or a clinician-examiner is used as a recorder.
SUMMARY

OBJECTIVE
Standardized patient-based examinations commonly use checklist recordings to evaluate students’ clinical performance. This paper examines whether and to what extent item and raters’ characteristics affect the reliability of history checklist recording of a standardized-patient-based assessment.

METHODS
Checklist items were reviewed for presence or absence of five item characteristics and 2- versus 3-point scoring scale. Agreement between checklists recordings obtained from standardized patients (SPs) and clinician-examiners (CEs) were compared by item characteristics, scoring scale and CEs’ level of involvement in the assessment.

RESULTS
Based on 3179 pairs of recording, the overall percentage of agreement between SPs and CEs was 83% (kappa = 0.64). Agreement was significantly higher for items scored on a 2-point than on a 3-point scale, and when the CE was also the author and the trainer of the station. After controlling for other factors, item characteristics were only marginally associated with level of interrater agreement.

CONCLUSIONS
This study suggests that attention should be paid to specific aspects of checklist development and checklist recording training when a standardized patient or clinician-examiner is used as recorder.
INTRODUCTION

Standardized patient-based examinations commonly use checklist recordings to evaluate students’ clinical performance. Depending on test feasibility, cost-efficiency, aims, or evaluation policy, testing programs rely on standardized patients’ (SPs), clinician-examiners’ (CEs) checklist recordings, or both, to derive evaluation of students’ clinical performance. Although SPs record various skills of students’ performance accurately,\(^1\)\(^-\)\(^4\) evidence suggests that SPs’ recording accuracy is affected by factors such as level of training in checklist completion, checklist length, and type and clarity of checklist items.\(^1\),\(^5\)-\(^7\) Clinician-examiners interrater reliability is generally considered to be high\(^8\), particularly when judgments are recorded on structured forms.\(^9\) The literature is more conflicting regarding whether CEs’ reliability is affected by their level of training\(^10\) and experience in assessment.\(^11\) A recent study suggests that interrater agreement is associated with examiner involvement in station construction but not with examiner years of experience in examining.\(^6\)

How to improve interrater reliability on checklist recording is also unclear. For instance, checklist development process is seldom reported in details; therefore, assessing the relations between aspects of checklist development and recording reliability is difficult.\(^12\) Although guidelines are available for writing good checklist items, they may not be consistently used in practice. Furthermore, little is known about how characteristics of the checklist items and type of checklist recorders interact to affect interrater agreement.

The present study explored: (a) whether some checklist item characteristics affected agreement between SPs and CEs on the history items of an end-of-clerkship SP-based examination, (b) whether relationships between item characteristics and interrater agreement in checklist recordings were related to the CEs’ level of involvement in the assessment, and finally (c) whether differences in SPs and CEs’ agreements affected students’ pass-fail status.
METHODS

Since 1995, the University of Geneva Faculty of Medicine had introduced a six-year integrated problem-based curriculum. An SP-based performance assessment takes place at the end of each internal medicine clerkship rotation in the fourth and fifth years. Students are instructed to conduct a 20-minutes focused patient history and physical examination based on the patient’s presenting chief complaint. All cases are derived from the list of presenting complaints commonly encountered in internal medicine. Students’ performance is recorded independently on checklists by the CEs and the SPs. While the SPs fill out the checklist at the end of the encounter, the CEs do so while observing the students.

This study used data collected during 9 consecutive examination sessions conducted from 1998 to 2000. Students’ ratings and pass-fail decisions were based on CEs’ checklists.

All the cases, their checklists, and scoring keys were developed and reviewed by a reference panel of eight faculty members of the department of Internal Medicine. The author of the case trained the SPs during two separate training sessions to portray the case and record the checklists. All SPs recruited for the examinations were experienced and have already functioned as SPs for two or three years.

Furthermore, they also have been verified and confirmed while functioning in previous teaching sessions for their reliability and accuracy in portraying various patient roles and in filling out checklists. Author-trainers also served sometimes—but not systematically—as CEs during the examination session; they were referred to as “trainer-examiners”. Other CEs were faculty members with teaching responsibilities during the clerkship and did not get directly involved in the elaboration of the SP-based assessments; they were referred to as “examiners”. Typically, examiners received the case scenario and the corresponding checklists a few days before the examination took place. They were coached to record the checklists by the case author through the witnessing of one of the two SP training sessions. Neither SPs nor examiners received further instruction between or during examination sessions.

Three potential sources of disagreement between SP and CE assessment of student performance were explored. Items on the history checklist were first classified according to whether they were scored on a
2-point- (asked – not asked) or a 3-point scale (completely asked - partially asked – not asked). Then, each checklist item was evaluated and classified in terms of the presence or absence of one of the following five characteristics perceived as likely to influence interrater agreement: (a) requirement that more than one piece of information be provided to receive a point (e.g., “the student asked me if I had nausea and if I vomited”); (b) use of medical terminology (e.g., “the student asked me if I had melena”); (c) no explicit indication on how to check the item (e.g., “the student asked me if I had fever or if I had shaking chills” versus “the student asked me if I had fever (1 point) or if I had shaking chills (1 point)”); (d) use of unclear or ambiguous wording or sentence; and (e) no explicit inclusion in the item of SPs’ expected answer to the students’ question (e.g., “the student asked me if I had fever or shaking chills” versus the same sentence with the following mention (I had fever but no shaking chills’”). The first two item characteristics have been previously described in the literature, while the other three have not been formally investigated until now.

Cohen’s Kappa statistic was used to compare percent of interrater agreement by item and rater characteristics. Comparison among Kappa statistics was based on the 95% confidence interval method. Interrater agreement was also studied with a random effects logistic model. In this model, items were nested within the students and adjusted for the item scoring system (2-point or 3-point), the number of history checklist items, the five item characteristics, and whether the examiner was a trainer. Bartlett’s test was applied to further analyze the interrater agreement by comparing variance between the scores given to the students by SPs, examiners and trainer-examiners. Exact McNemar test was used to determine to what extent differences in SPs and CEs’ agreements affected students’ pass-fail decisions.

RESULTS

We reviewed performance evaluation of 229 students on a total of 36 cases (mean: 6 evaluations per case). Twenty cases were used once, and 8 twice with minor checklist revisions on the second time.
Cases were portrayed by 30 different SPs. Five SPs played the same role twice with an interval of 9 to 17 months between the two examination sessions. Twenty-six CEs were involved (18 “examiners” and 8 “trainer-examiners”). Mean number of items per checklist was 33 (SD = 3.4; range = 24 to 41). Each checklist included an average of 14 history items (SD = 1.98; range = 10 to 18), 12 physical-examination items, and 7 items on communication and interpersonal relationship. Further analyses were restricted to history items (n= 3179). Of them, 80% were coded on a 2-point-scale (n= 2535) and 20% (n= 644) on a 3-point-scale; 64% had none of the five studied item characteristics, and 36% contained from one to a maximum of 4 characteristics. One fourth of items (25.3%) required that more than one piece of information be provided to receive a point; 6.1% used medical terminology; 14.4% did not include explicit indication for item checking; 8.1% had unclear wording; and 7.3% did no list answers expected from SPs.

Percentage of agreement between SPs and CEs on history items was substantial (83% with a corresponding kappa coefficient of 0.64, 95% CI: 0.61-0.67). In univariate analysis (see Table 1), percentage of agreement was 11.6% significantly higher for items coded on a 2-point scale (kappa: 0.66, 95% CI: 0.63-0.69) than for items coded on a 3-point scale (kappa: 0.53, 95% CI: 0.47-0.59; P< 0.05).

Interrater agreement tended to be consistently higher for items without the studied characteristic than for items with that characteristic, but difference in agreement reached significance only for the item characteristic “require more than one piece of information” (+ 9.4%; kappa for items without characteristic: 0.66, 95% CI: 0.63-0.70 vs. kappa for items with characteristic: 0.56, 95% CI: 0.51-0.61; P<0.05).

Overall agreement between SPs and CEs was 4.5% significantly higher when the CE were trainer-examiners than when the CE were examiners (kappa: 0.70, 95% CI: 0.66-0.75 vs. kappa: 0.62, 95% CI: 0.58-0.65; P<0.05). In stratified analysis, this higher level of agreement between SPs and trainer-examiners than between SPs and examiners was only significant for items which have none of the five studied characteristics (+7.7%; kappa: 0.69, 95% CI: 0.62-0.77 vs. kappa: 0.56, 95% CI: 0.50-0.61;
P<0.05) but not for those which have one or more of those characteristics (+2.8%; kappa: 0.71, 95% CI: 0.65-0.78 vs. kappa: 0.65, 95% CI: 0.61-0.69).

Variance analyses were conducted to determine whether the observed differences between examiners and trainer-examiners could result from the variability among the SPs or the CEs. Overall, the mean scores of history checklists given by SPs were higher than the mean scores given by CEs (paired T test: 14.76 ± 4.35 vs. 14.0 ± 4.45, P =0.00001), and the variance heterogeneity among SPs was similar to the one among the CEs (variance comparison test according to Leven, Brown and Forsy: P = 0.61). The Bartlett’s test further showed that the score variability was not different between the group of SPs working with examiners (SD = 4.43) and the group of SPs working with trainer-examiners (SD = 3.83, P =0.17). However, the score variability was significantly higher among examiners (SD = 4.61) than among trainer-examiners (SD = 3.58, P =0.02).

SPs and CEs disagreed on pass-fail decisions for 16.6% (n= 38/229) of the students. If only SPs checklists had been used, 7% (n=16) more students would have passed the exam (Exact McNemar test: P =0.01). Disagreement on pass-fail decisions between SPs and examiners tended to be higher (19.5%; kappa: 0.57, 95% CI: 0.44-0.70) than disagreement between SPs and trainer-examiners (9.2%; kappa: 0.76, 95% CI: 0.58-0.94), but this difference was not statistically significant.

In multivariate analyses (Table 2), two variables independently predicted a higher level of agreement between SPs and CEs: item coded on a 2-point- vs. 3-point scale (P=0.01) and CE referred as trainer-examiners vs. examiners (P= 0.007). Agreement between SPs and CEs was marginally lower for items using medical terminology than for items not using such terminology (P=0.07). Number of items per history checklist was also marginally associated with level of agreement (P=0.07), but with an odds ratio close to 1.0 (OR: 1.1) and a very narrow 95% CI (0.99-1.1).
DISCUSSION

Our results suggest or confirm the approaches by which we can increase the agreement between SPs and CEs on history checklists. Although overall agreement between SPs and CEs in this study is comparable to what has been reported elsewhere,¹,³ agreement is independently associated with CEs’ level of involvement in the assessment, and with checklist item type of coding.

Univariate percentage agreement, kappa statistics, and multivariate analyses concur to indicate that interrater agreement is lower for items coded on a 3-point- compared to a 2-point scale. Whereas the chance probability of interrater agreement on a 3-point-scale (33%) is lower than on a 2-point-scale (50%), the Cohen’s Kappa coefficient which controls for this source of variability indicate that the observed differences in agreement do not result from chance alone. According to the literature, interrater agreement is lower for physical-examination items testing for several steps of a procedure and for history items that include more than one piece of historical information.¹ Hence, we may have found a higher level of discordance for items rated on a 3-point scale in this study because these items require multiple observations to give the mark and are probably more difficult to assess when there is uncertainty about student’s performance. In our study, items coded on a 3-point-scale were mainly used to look for two related pieces of historical information with only one item (e.g., “the student asked me if I had nausea and if I vomited”) in an attempt to reduce the number of checklist items. Our data suggest that interrater agreement could be better preserved by keeping the items separated rather than by combining them. Although there is concern that increasing the length of a checklist might negatively affect interrater agreement,¹ the maximum acceptable number of items in a checklist has not been precisely determined. In this study, checklist length seems not to be related to interrater agreement (table 2) and substantial agreement between SPs and CEs is achieved on history items (83%) despite an overall exam checklist longer (33 items) than what is usually recommended.¹
The other four items characteristics, which were considered a priori as weaknesses, are found to be not, or only marginally, associated with lower interrater agreement. Two of the characteristics focused, in fact, on the absence of built-in cues to improve SP's recording accuracy (providing an explicit indication on how to record the item and providing a reminder of the SPs’ expected answer). Although we find no evidence that these cues are useful, it would be premature to conclude they should be abandoned. These characteristics deserve to be further addressed with a controlled prospective study to determine, for example, whether they are more useful for certain types of history items than others.

Unclear wording and the use of medical terminology must be avoided for obvious reasons. In our study, these characteristics are not associated with lower interrater agreement, but possible explanation for this is that ambiguities were clarified verbally during the training sessions or that these items similarly influenced both the SPs and the CEs.

Though the observation does not extend to pass-fail decisions, an important finding is that overall agreement between SPs and CEs is higher with trainer-examiners than with examiners. At our institution, trainer-examiners authored the cases. They also spent two sessions training the SPs to portray the cases and record the checklists, whereas examiners discussed ambiguities in checklist item formulation and recording during only one training session. Finally, compared to trainer-examiners, examiners’ involvement in assessment was limited. This suggests, as previously showed at our institution\textsuperscript{11} and elsewhere,\textsuperscript{6,10} that CEs’ involvement in station construction, and CEs’ training and experience in assessment may have a positive impact on interrater agreement. Can the differences observed between trainer-examiners and examiners result from the variability attributable to the SPs? The fact that the score variability is significantly higher for examiners than for trainer-examiners and that there are no differences observed in the SPs’ score variability further reinforced the observation that the difference in interrater agreement is more likely to be related to CEs characteristics than to those of the SPs.
Our study has several limitations. First, the design was observational, with no “gold standard” measures against which to assess SPs’ and CEs’ recording accuracy. Second, SPs and CEs were in the same room during the examinations and communications between them - the so called contamination bias - may have occurred despite instruction not to do so. The logical consequence of discussing checklist recording, however, would be to increase interrater agreement; therefore had SPs and CEs exchanged opinions about recording, the impact on the positive findings of this study should be minimal. Third, because the number of SPs and CEs involved in the study was large, we were unable to estimate the performances of each SP and CE separately.

In conclusion, history checklists should avoid, when possible, items with a 3-point scale coding, since interrater agreement seems to be better preserved by coding on a 2-point scale even at the cost of increasing the number of items. The level of involvement of CEs in station construction or their level of experience in assessment also influences interrater agreement especially for flawed items. Further studies are needed to determine the relative impact on interrater agreement of the CEs’ involvement in station construction, level of participation in training session and level of experience in assessment.
REFERENCES


<table>
<thead>
<tr>
<th>Items</th>
<th>n</th>
<th>Agreement (%</th>
<th>Kappa *</th>
<th>95 % CI in agreement</th>
<th>Absolute difference in agreement</th>
<th>P-value **</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Response Scoring</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Items scored on a 2-point scale</td>
<td>2535</td>
<td>85.4</td>
<td>0.66</td>
<td>0.63-0.69</td>
<td>+ 11.6%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Items scored on a 3-point scale</td>
<td>644</td>
<td>73.8</td>
<td>0.53</td>
<td>0.47-0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Item characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Require more than one piece of information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>2376</td>
<td>85.4</td>
<td>0.66</td>
<td>0.63-0.70</td>
<td>+ 9.4%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Present</td>
<td>803</td>
<td>76.0</td>
<td>0.56</td>
<td>0.51-0.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No explicit indication on how to check</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>2720</td>
<td>84.1</td>
<td>0.65</td>
<td>0.62-0.68</td>
<td>+ 7.6%</td>
<td>NS</td>
</tr>
<tr>
<td>Present</td>
<td>459</td>
<td>76.5</td>
<td>0.56</td>
<td>0.49-0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclear wording</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>2921</td>
<td>83.5</td>
<td>0.65</td>
<td>0.62-0.67</td>
<td>+ 6.4%</td>
<td>NS</td>
</tr>
<tr>
<td>Present</td>
<td>258</td>
<td>77.1</td>
<td>0.58</td>
<td>0.49-0.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use medical terminology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>2985</td>
<td>83.4</td>
<td>0.65</td>
<td>0.62-0.67</td>
<td>+6.6%</td>
<td>NS</td>
</tr>
<tr>
<td>Present</td>
<td>194</td>
<td>76.8</td>
<td>0.55</td>
<td>0.45-0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPs’ expected answer are not mentioned</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>2947</td>
<td>83.5</td>
<td>0.64</td>
<td>0.62-0.67</td>
<td>+ 6.3%</td>
<td>NS</td>
</tr>
<tr>
<td>Present</td>
<td>232</td>
<td>77.2</td>
<td>0.59</td>
<td>0.49-0.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Items with none of the above 5 characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Items with one or more of the above 5 characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>2049</td>
<td>85.5</td>
<td>0.66</td>
<td>0.63-0.7</td>
<td>+ 7%</td>
<td>NS</td>
</tr>
<tr>
<td>Present</td>
<td>1130</td>
<td>78.5</td>
<td>0.59</td>
<td>0.55-0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Type of examiners</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trainer-examiners</td>
<td>852</td>
<td>86.3</td>
<td>0.70</td>
<td>0.66-0.75</td>
<td>+ 4.5%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Examiners</td>
<td>2327</td>
<td>81.8</td>
<td>0.62</td>
<td>0.58-0.65</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Cohen’s Kappa statistic.

** Comparison among Kappa was based on the 95% confidence interval method.

NS = non significant.
Table 2. Random effects logistic model predicting interrater agreement between standardized patients (SPs) and clinician-examiners. (n=3179)

<table>
<thead>
<tr>
<th></th>
<th>Odds Ratio</th>
<th>Std. Error</th>
<th>z</th>
<th>p</th>
<th>95 % CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trainer-examiners (vs. examiners)</td>
<td>1.46</td>
<td>0.20</td>
<td>2.69</td>
<td>0.007</td>
<td>1.11-1.91</td>
</tr>
<tr>
<td>Coding on a 3 point scale (vs 2 points)</td>
<td>0.55</td>
<td>0.13</td>
<td>-2.48</td>
<td>0.01</td>
<td>0.35-0.88</td>
</tr>
<tr>
<td>Use medical terminology</td>
<td>0.71</td>
<td>0.13</td>
<td>-1.83</td>
<td>0.07</td>
<td>0.49-1.03</td>
</tr>
<tr>
<td>Number of items per history checklist</td>
<td>1.05</td>
<td>0.03</td>
<td>1.8</td>
<td>0.07</td>
<td>0.99-1.12</td>
</tr>
<tr>
<td>Unclear wording</td>
<td>0.81</td>
<td>0.14</td>
<td>-1.16</td>
<td>0.24</td>
<td>0.58-1.15</td>
</tr>
<tr>
<td>SPs’ expected answer are not mentioned</td>
<td>0.83</td>
<td>0.16</td>
<td>-0.94</td>
<td>0.35</td>
<td>0.57-1.22</td>
</tr>
<tr>
<td>No explicit indication on how to check</td>
<td>0.95</td>
<td>0.17</td>
<td>-0.30</td>
<td>0.77</td>
<td>0.67-1.34</td>
</tr>
<tr>
<td>Require more than one piece of information</td>
<td>1.00</td>
<td>0.26</td>
<td>-0.01</td>
<td>0.99</td>
<td>0.60-1.65</td>
</tr>
</tbody>
</table>

Note: Items are nested within students while adjusting for items characteristics, coding system, type of examiners and number of items per history checklist