Supplement B: Data validation

Programmatically, data size, format, consistency and range were validated, batch total and logic checks (an item can be either correctly, incorrectly or unclearly performed, item dependencies should correspond with their ratings) performed.
Quantile-quantile plots are shown in figure 7. No normal distribution is discernible, supporting the choice of non-parametric tests. Tests for distribution were deemed superfluous hereafter.
To stress the validity of the data used for analysis, the plausibility of rater disagreement on performance is investigated hereafter. The authors hope that this investigation may guide methodical improvements of subsequent studies.

Agreement Outlier Discussion

Because marginal totals for agreement on performance are imbalanced in the corresponding contingency table, Cohen's Kappa (sensitive to this imbalance to the point of paradoxical behavior) was not calculated. Uni-directional proportions of agreement (p_{pos}, p_{neg}), their sum weighted by proportions of total ratings (p_0) and prevalence- and bias-adjusted kappa (PABAK, $K = 2p_0 - 1$) were calculated per team and per item, instead (see figure 6).
Outlying teams: Minima in proportionate positive agreement on performance were identified for one supported and one unsupported team of non-professionals ($p_{pos} = .48$ and $p_{pos} = .40$) and for one team of unsupported professionals ($p_{pos} = .36$). No other proportionate positive agreement on team performance undercut $p_{pos} = .61$.
The team of unsupported non-professionals spent nearly eight minutes trying to survey blood pressure, pulse rate and pupillary light reflex, then decided they had finished the task. This leaves very few items for the raters to agree on, some of which were repeatedly performed but not consistently correctly. The team of unsupported professionals decided they had completed the task after less than three minutes, not actually performing most of the surveys. This misunderstanding by participants, whether and to what extent activities were expected to be performed in a simulation setting, opened room for rater disagreement. The team of supported non-professionals, while performing most items, left out arguably critical steps such as in asking for the last meal not asking for the exact time. As another example, in measuring the breathing rate, they counted for too short a time, straying from the correct value by an arguably relevant amount. This room for interpretation may have been particularly large during the first recordings to be rated, when clarity of the frame of reference was only beginning to evolve.
Items of disagreement: Among professionals, proportionate positive agreement on items 3-6, 8, 16 and 17 undercut $p_{pos} = .63$.
Among non-professionals, that on items 4, 6, 8, 10, 11, 16 and 17 undercut the same threshold.

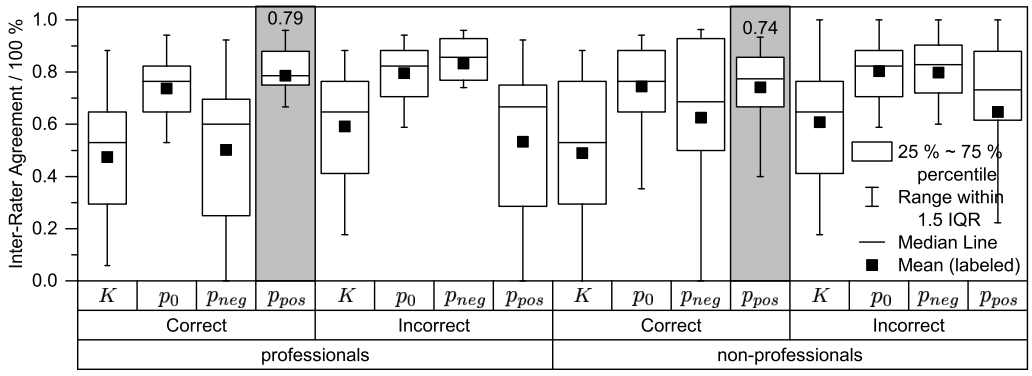

Figure 6: Overview of inter-rater agreement on skewed data per team. Inter-rater agreement on items rated correctly or incorrectly performed by professionals or non-professionals: prevalence- and bias-adjusted kappa (PABAK, K), the proportion of observed agreement p_0, which is the sum of observed proportion of positive agreement p_{pos} and observed proportion of negative agreement p_{neg} weighted by positive and negative ratings per reviewer and team. In a range of $[0, 1]$, higher means more inter-rater agreement. Highlighted and labeled agreement dimensions characterize score dimensions used for quantitative analysis.

Item 3 concerns the Glasgow-Coma-Scale. No team of unsupported non-professionals but all teams of unsupported professionals at least attempted to survey this item. In contrast, all supported teams attempted to survey it. Thus, the room for disagreement on correct performance was reduced for non-professionals. The interpretation of correct performance seems to have ranged from accepting an intuitive close-enough estimation to accepting only accurately deduced scores, hence the low agreement.
Item 4 concerns normal breathing. One rater rated normal breathing and rate (item 5) independently with the rate being more often correctly surveyed:

- Items 4 and 5 were differently rated in 7 out of 15 teams who correctly performed either.
- In 6 of these 7, item 5 was rated as correctly performed without item 4 being rated as correctly performed.

The other rater rated "normal" as more often correctly performed:

- Both items were rated differently in 4 out of 6 teams who correctly performed either.
- Only in 1 out of these 4 teams was item 5 rated as correctly performed without item 4 having been rated so.

Conclusions should not be drawn from this tiny sample, a difference in interpretation of "normal" among raters seems to be hinted at.
Item 5 concerns the breathing rate. Raters agreed less on teams of professionals, of whom nearly twice as many surveyed the item at all compared with non-professionals. Among this higher number of potentially correctly performed items, the aforementioned difference in interpretation of arguably relevant deviation from the simulated rate may be dominant. Closely related is *item 6 concerning correctness of respiratory rate measurement*: One rater rated the measurement correctly performed in 14, the other in only 5 teams. The authors hypothesize that the raters prioritized minimal measurement duration differently.
In item 8 on correctly surveying radial pulse, like in item 6, one rater rated the survey correctly performed in 32, the other in only 11 teams. The authors hypothesize that the correct result of the measurement was for only one rater integral to the item. However, this result might not have been attained once the rate was available from the ECG in intervention groups. Supported were only 2 out of the 11 teams who correctly performed palpation. Of all 18 supported teams, the same rater rated 13 as having tested for radial and carotid pulse (item 7). If the result from palpation were not integral to the correct survey, 11 supported teams of non-professionals and professionals would have been rated as not having correctly performed palpation.
On items 10 and 11 concerning blood pressure and whether it is normal, proportionate positive rater agreement on non-professionals was low. Particularly, raters disagreed on correct performance by 6 supported and 3 unsupported teams. In 5 of the supported teams, the monitoring device measured a blood pressure lower than the simulator was programmed to output and the threshold criterion for rating the item correctly performed. One rater therefore did not mark these items correctly performed. The remaining supported team never notified the remote physician of the blood pressure, which one rater therefore deemed not measured, not knowing that the remote physician had received the measurement result from the monitoring device. Among the unsupported teams, one simulator malfunction, one supposition on normal blood pressure without measurement and one case of incorrectly performed item 9 (correctness of measurement) created uncertainty for the raters.
Item 16 concerns asking about the last food intake. One rater rated it as incorrectly performed 10 times more often than the other, each time because the exact time of the last meal had not been inquired after. Only the general time of day had been asked after, which the other rater deemed sufficient.
Item 17 concerning the course of accident was rated differently for 24 teams. The rater who rated the item these 24 times incorrectly performed noted down for 22 of these teams that the exact height of the fall of the patient had not been explicitly asked after. In 8 of these 24 teams, the burn that caused the fall had been left out of the course of accident. The other rater rated less accurate courses of accident as correct item performance in all of these 24 teams.
In conclusion, the rater disagreement seems to stem from

1. simulation artifacts causing participants to behave differently in an artificial setting than they might in reality,
2. an emerging frame of reference in which the

raters' interpretations became consistent,

3. differences in raters' prioritization of accuracy of deductions versus precision of adherence to instructions or guidelines,
4. differences in raters' understanding of item dependencies.

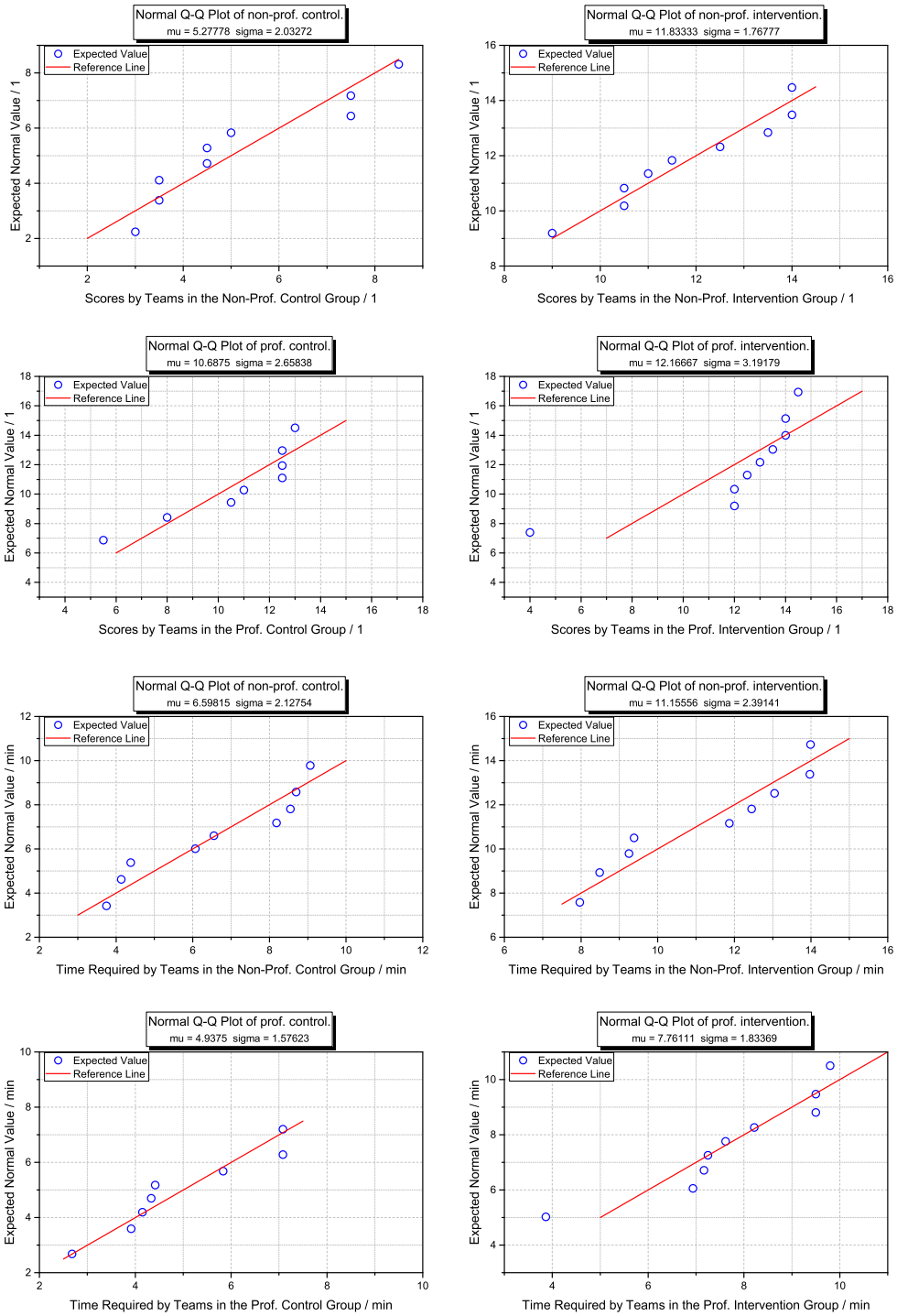

Figure 7: Quantile-quantile plots for every investigated group.