BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

Appendix B

A patient's performance on a single collection is measured as the percentage of correct responses and a response is evaluated as correct if no errors were made as well as no help button was used.

In a next step, the guessing probability is calculated by considering permutation when appropriate. If the guessing probability does not exceed a predefined threshold (0.2) data will be analysed based on the sequential probability ratio test (SPRT),[47] otherwise by using confidence intervals and predefined thresholds (0.7 and 0.8). The outcome of these analyses indicates how the difficulty level should be adjusted and the algorithm assigns a new collection consisting of exercises with one more distractor if the difficulty level is to be increased, for example.

It is crucial to note, according to the SPRT that the lower the guessing probability is, the fewer correct responses yield a significant difference (e.g. the easiest is a raise of difficulty level). To ensure a moderate increase of difficulty level, a threshold of 0.2 was set as minimal guessing probability.

In a following step the type of binomial distribution which fits the data best is analysed, e.g. two models are postulated and compared to each other (Figure 6). The first model (h_0), constituting the null hypothesis, describes a binomial distribution with the guessing probability set as the probability parameter p_1 of the distribution. The second model (h_1) defines the alternative hypothesis by postulating a binomial distribution with a probability of 0.9 which is regarded as a sufficient performance to raise exercises' difficulty level.

To decide how to adjust the difficulty, the logarithmic ratio of the model constituting the alternative hypothesis to the other model is compared with criteria which are based on the chosen alpha and beta level. A logarithmic ratio of greater than 0 indicates that the model realising the alternative hypothesis fits the data better than the model with guessing probability.

To evaluate whether a difference of fit is significant or not, alpha and beta were set to the stringent .001 level. This level of significance was used to ensure that the decision to stay at the current difficulty level was somewhat probable. This becomes plausible, if we take into consideration the inverse proportionality between significance level and the probability staying at the current difficulty level which is valid for the implemented SPRT. If a binomial model with a probability of 0.9 fits the data significantly better than the model with guessing probability than the difficulty of the next collection gets raised and if it fits the data significantly worse than the difficulty gets lowered. In order to take the stringent evaluation of responses into account (a response is correct if no errors were made and the help button was not used), a non-significant finding can heighten difficulty if the value of the logarithmic ratio is greater than 0.

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

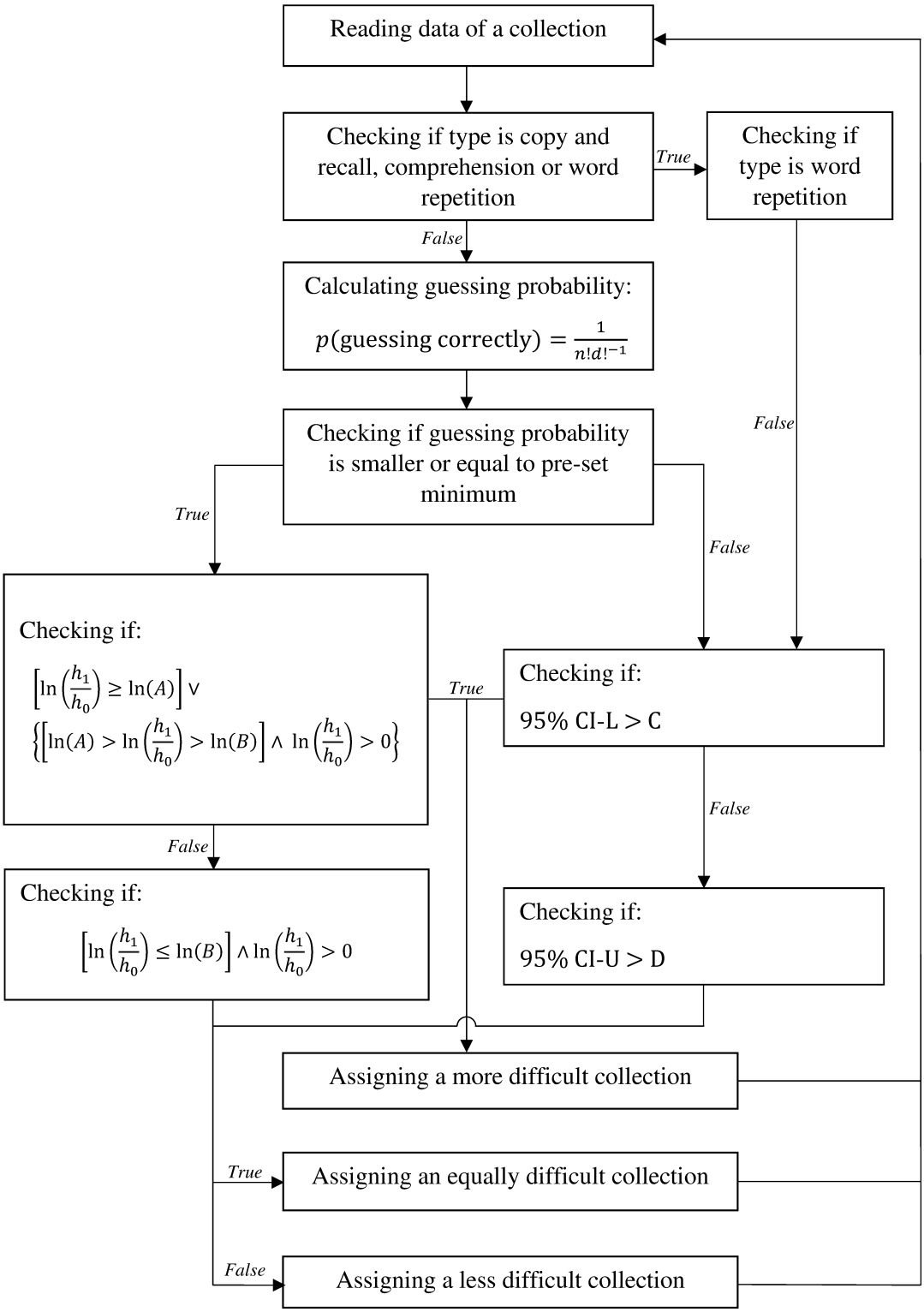

Figure 6. Detailed flowchart of the computerised adaptive testing algorithm with n = number of response elements, d = number of distractors, $h_i = \binom{n_i}{k_i} {p_i}^{k_i}(1-p_i)^{n_i-k_i}$, $A = \frac{1-\beta}{\alpha}$, $B = \frac{\beta}{1-\alpha}$, *upper and lower boundaries of confidence intervals (CI-U = 0.8, CI-L = 0.7).*

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

If the SPRT is not appropriate, data are analysed with confidence intervals and thresholds. Rather than taking the recent performance (e.g. the mean performance of the last 25 exercises), all data collected within a timeframe are considered in order to minimise error variance. Therefore, a 95% confidence interval around the mean performance is calculated in order to detect significant differences from predefined thresholds. If the difference is significant, the difficulty level gets adjusted (raised or lowered) and otherwise an equally difficult collection is assigned. Rather than a descriptive approach, this inferential statistical analysis was implemented to fix the alpha error rate. Otherwise, it would remain unclear how probable an erroneously increased difficulty level is which in turn impairs rehabilitation progress. To increase the difficulty of exercises, the lower bound of the confidence interval must be greater than 0.8. Whereas if the upper bound of the confidence interval is smaller than 0.7 than difficulty gets reduced.