Online Appendix: Measuring Student Engagement During Collaboration
Interpretation of response kernel of the Hawkes process

The response functions for the multidimensional Hawkes process can be written as

$$\phi_{jk}(u) = \alpha_{jk} f_{jk}(u)$$

where $f_{jk}$ is a density on $\mathbb{R}^+$, which we refer to as the response kernel, and $\alpha_{jk} \in \mathbb{R}^+$ is referred to as the response intensity. Notation is described in Equations 4 through 7 of the main paper, and the response intensity parameter is discussed in Equations 9 through 12 of the main paper. Here we discuss the response kernel.

Following an action of a student during a CPS task, the response kernel can be used to characterize various qualities of the immediate future such as: (a) the latency period before the further actions are expected from other team members, (b) the period of time over which further actions are most likely, and (c) the rate of decay in the probability of further actions. As noted by Hawkes (1971), the mathematics of the process are especially tractable when the response kernel is chosen to be the exponential distribution, and Liniger (2009) suggested that the exponential kernel be used whenever there is no strong preference. This choice also has been a prominent feature of many statistical applications, notably in seismology (e.g., Ogata, 1988). However, in application to CPS, choosing a flexible family of response kernels is important to accommodate differences in the dynamics of individual team members, across different teams, across different tasks, and for different modalities of interaction (e.g., chat, email, face-to-face). Halpin and De Boeck (2013) suggested the use of the two-parameter gamma density for the response kernel, and found this to provide reasonable fit for email transactions. Halpin and von Davier (n.d.) also found acceptable fit of the gamma kernel in application to teamwork in sports. In the present research we
apply the gamma kernel to online chat during a CPS task.

Additional comments on maximum likelihood estimation

As mentioned in main paper, recent research on estimation of the Hawkes process has addressed applications of the EM algorithm using the branching structure representation. Initial work was motivated by concerns about the curvature of the likelihood function at the MLEs (see Veen & Schoenberg, 2008). In particular, the log-likelihood of the Hawkes process contains a weighted sum of density functions, leading to challenges in numerical optimization that are familiar from research on finite mixture modeling (e.g., McLachlan & Peel, 2000) and non-linear regression (e.g., Seber & Wild, 2003). Definitive recommendations about the required sample size of such models are not currently available, although past work has found that the EM algorithm led to acceptable conditioning of the bivariate Hawkes process with $n_j \approx 200$ (Halpin & De Boeck, 2013). In the empirical example used in the present research, the number of events is even smaller, which led us to consider restrictions on the response kernels (see following sections).

While numerical methods can be used to estimate the asymptotic covariance matrix of the MLEs of the Hawkes process using standard results (Ogata, 1978), closed-form results have only been obtained in special cases (e.g., Ozaki, 1979). This is an unfortunate limitation of current research, because analytic expressions are important for understanding how precision of estimation can be controlled via the design of CPS tasks. To address this point, we now provide an informal argument leading to see Equation 13 of the main paper:

$$V[\hat{\alpha}_{jk}] > \frac{\alpha_{jk}}{\sum_{i=1}^{n_k} F_{jk}(T - t_{ik})}.$$
The argument is based on the complete-data log-likelihood of the Hawkes process, derived via its branching structure representation (see Halpin & De Boeck, 2013). It relies on two basic observations: (a) the standard errors are easily obtained using the complete data log-likelihood and (b) the missing information principle relates the Fisher information of the complete-data and incomplete-data log-likelihood functions, implying that standard errors of the former are a lower bound on those of the latter (see Meng & Rubin, 1991; Mclachlan & Krishnan, 2008).

Let $\tilde{\alpha}_{jk}$ denote estimates of $\alpha_{jk}$ obtained by maximizing the complete-data log-likelihood. Halpin and De Boeck (2013, Eq. A11) showed that

$$\tilde{\alpha}_{jk} = \frac{\sum_{i=1}^{n_k} n_{ijk}}{\sum_{r=1}^{n_k} F_{jk}(T - t_{rk})},$$

in which the number of events $n_{ijk} \geq 0$ generated by the Poisson process $N_{jk}(t_{ik}, T]$ is assumed to be observed.

It is important to note that the denominator of (3) is constant with respect to expectations taken over the processes $N_{jk}(t_{ik}, T]$, when $j \neq k$. To see this, recall that the notation $N_{jk}(t_{ik}, T]$ denotes the Poisson process governing responses of student $j$ to action $i$ of student $k$. More specifically, the first subscript on $N_{jk}$ denotes the student whose actions are being modeled via the Poisson process. Thus it is apparent that, when $j \neq k$, $t_{ik}$ cannot be due to any process $N_{jk}$ – this is merely a matter of notation.

Consequently, we require only the variance of the numerator of Equation (3). By definition of the branching structure, each $n_{ijk}$ is given by a Poisson distribution with rate $\alpha_{jk} F_{jk}(T - t_{ik})$, and the Poisson processes that make up the branching structure are inde-
V[\hat{\alpha}_{jk}] \geq V[\tilde{\alpha}_{jk}], where \hat{\alpha}_{jk} is the MLE based on the incomplete-data log-likelihood. We leave a more thorough characterization of the sampling behavior the Hawkes process to further research.

Additional details on the Tetralogue task and sample data

Our example was obtained from the Tetralogue, a two-player simulation-based science game with an embedded assessment that has been recently developed by Educational Testing Service. Like the Assessment and Teaching of 21st Century Skills program (Griffin, McGaw, & Care, 2012; Griffin & Care, 2015), the purpose of the Tetralogue is to jointly assess cognitive (science) skills and social (CPS) skills. The Tetralogue was developed from the Trialogue, a single-player version in which a student interacts with two computer agents during the simulation (Zapata-Rivera et al., 2014; Zapata-Rivera, Liu, Chen, Hao, & von Davier, n.d.). The addition of a second student, with interactions between students accomplished via online chat, provided the technology platform of the Tetralogue.

During the simulation, dyads work together to learn and make predictions about volcano activity. At various points in the simulation, the students are asked to individually submit their responses to an assessment item without discussing the item. Following submission of responses from both students, they are invited to discuss the question and their answers. Lastly, they are given an opportunity to revise their responses to the item, with the final answers counting towards the team’s score. In addition to the computer agents, a
system prompt explicitly instructs students when to share, explain, evaluate, compare, and contrast their individual knowledge, which is intended to scaffold appropriate interactions (e.g., Rummel & Spada, 2005; Shaffer & Gee, 2012; Walker, Rummel, & Koedinger, 2011). The system prompt and the computer agents are specifically designed to elicit actions that fall under the Tetralogue conceptual model (Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015). This conceptual model describes the individual and social skills required for successful task completion, and is currently being adapted to the development of further CPS tasks.

The dataset used for the present analysis contained a total of 286 dyads. Individual participants were solicited using the crowdsourcing service Amazon Mechanical Turk and were paired with one another based on their arrival in the queue. A total of 81.5% of the sample responded to a demographic survey. The median reported age was 31.5 years, 52.5% reported that they were female, and 79.2% reported that they were White. Additionally, all participants were required to (a) have an IP address located in the United States, (b) self-identify as speaking English as their primary language, and (c) self-identify as having at least one year of college education. The sample and measures are described further by Hao et al. (2015).

The recorded time-to-completion of the 286 dyads was markedly bimodal, with 268 teams spending less than 130 minutes on the task, but the remaining 18 teams spending over 800 minutes. While the latter cases were presumably due to a recording error, it was not apparent what the caused this or how could be adjusted, so data from these dyads were not analyzed. Figure 1 plots the number of chats as a function of time-to-completion for the remaining teams. The range of total of chat messages per dyad was [37, 232]. For the individual participants, the range of chat messages sent was [17, 126]. Consequently, for many
dyads, the number of chat messages sent between partners was too small to support time series modeling. To address this issue, we omitted dyads whose total number of messages was less than 85, and, as described below, we also fitted a single response kernel to all of the chat messages sent by a given dyad. The exact value of this cut-off was admittedly arbitrary, although we found that fewer chat messages frequently resulted in non-convergence of the estimation algorithm. As noted in the comments on estimation, further research is needed on the use of small sample sizes with the Hawkes process. The present example does not seek to address these issues, but merely serves to illustrate the methodology in application to CPS. The following analyses are reported for only the 90 remaining dyads.

![Figure 1](image)

*Figure 1.* For each team, the number of messages sent as a function of number of minutes spent on task.

**Fitting the Hawkes process**

For each dyad, we first assessed whether the chat data demonstrated clustering, which, as described in main paper, is a basic assumption of the Hawkes process. A temporal
point process is defined to be clustered if the rate of events is overdispersed relative to the homogeneous Poisson process. In more practical terms, this means that the waiting times between some chat messages should be relatively long, but for most messages the waiting times should be much shorter. To test for clustering, we used an approach motivated by the so-called “time-change theorem.” The theorem states that the waiting times of the residuals of a correctly specified point process are exponentially distributed with rate of one (see, e.g., Daley & Vera-Jones, 2003, chap. 7). For each of the 90 dyads, we assessed the presence of clustering using qq-plots and Kolmogorov-Smirnov (KS) tests of the residual waiting times of the homogenous Poisson process against their theoretical exponential distribution.

Because the Tetralogue was designed to ask students to complete parts of the task without discussion, it naturally created relatively long pauses between clusters of chatting. However, in several cases we found that clustering was not distinctly present in the chat data. To deal with these cases, we added a random time “buffer” to simulate the independent work sessions. This procedure involved right-shifting all event times that occurred subsequent to the onset of a scheduled independent work session. Letting $C$ denote a random draw from an exponential distribution with rate parameter equal to the 95-th percentile of the dyad’s empirical waiting time distribution, and letting $S$ denote the time of onset of an independent work session, the time buffer was implemented by replacing all $t_{ij} > S$ with $u_{ij} = t_{ij} + C$. This procedure is clearly ad hoc, but it allowed us to make use of the chat data of all 90 dyads.

We then fitted the Hawkes process to chat data from each dyad using the EM algorithm described in Halpin and De Boeck (2013) and Halpin (2013). We fitted the model with $\phi_{jk}(u) = \alpha_{jk} f(u; \xi_{jk})$, where $f$ was chosen to be the two-parameter gamma density. This choice was made after exploratory analyses with an exponential kernel showed unsat-
isfactory model fit. To address the small number of chats per dyad, we required $\xi_{ij} = \xi$, which substantially reduced the number of parameters to be estimated while still resulting in acceptable fit for most dyads. Standard errors for parameter estimates were obtained using the Hessian of the log-likelihood evaluated at the MLEs. All analyses were carried out in R with code available from the first author.

The goodness of fit of Hawkes process for each dyad was assessed using the same approach described for the homogeneous Poisson process. Qualitative judgments about fit led to rejection of the Hawkes process for at least one partner in a total of 16 dyads. Of these, 12 dyads had a total number of chat messages in the range $[85, 120]$, which suggests that the problem may have arisen due to a small number of chats. There was no apparent explanation for the poor fit of the remaining dyads. We speculate that allowing the response kernels to vary over individuals would appreciably improve the fit of the model in all cases, although in the present sample the number of chats was not sufficient to assess this conjecture. We omitted from subsequent analyses, including those analyses reported in the main paper, all 16 dyads for whom the Hawkes process did not fit their chat data.

Limitations of the embedded assessment

There were many challenges with the embedded assessment. Here we summarize the sample data and explain our rational for focusing on item revisions as a measure of task performance. In the first place, item calibration of initial responses using a two-parameter logistic model with the full dataset revealed that all but one of the items were quite easy, and that four of the items had discrimination parameters less than one. Thus, although Hao et al (2014) reported a small pre- to post-revision gain when averaging over the entire sample,
it was infeasible to reliably estimate gains at the individual or dyad level.

Secondly, it was found that not all participants improved their scores after having the opportunity to revise their answers. Indeed, the majority (57.5%) of individuals in our final sample did not make any revisions to their responses after discussion with their partner, and of these, 51% had all six of the easier items correct on their initial response. Moreover, about 15.7% of the sample made a revision in the wrong direction – they revised at least one correct answer to be incorrect. Nearly half of these individuals also made at least one other revision in the correct direction. This situation made it nontrivial to define “successful performance” on the revised items at the individual level: Individuals who did not revise their answers may have had the correct answer on their initial response, and individuals who revised their answers may have made a revision in the wrong direction.

Given these challenges with the embedded assessment, we focussed simply on whether or not each participant revised at least one response after having the opportunity to discuss the question with his / her partner. This aspect of task performance is unambiguous and is plausibly related to chat engagement between partners. We grouped individuals into “Revisions” versus “No Revisions,” and considered the mean engagement within each of the two groups. As per the above, it is important to keep in mind that these two groups do not correspond to higher or lower scores on the revised item responses. The results are summarized in Table 1 below and Figure 2 of the main paper.
Table 1

Summary of group differences.

<table>
<thead>
<tr>
<th>Index</th>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
<th>Hedges' g</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>No Revisions</td>
<td>0.31</td>
<td>0.13</td>
<td>82</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Alpha</td>
<td>Revisions</td>
<td>0.36</td>
<td>0.10</td>
<td>66</td>
<td>0.40</td>
<td>0.20</td>
</tr>
<tr>
<td>Partner’s Alpha</td>
<td>No Revisions</td>
<td>0.31</td>
<td>0.14</td>
<td>82</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Partner’s Alpha</td>
<td>Revisions</td>
<td>0.37</td>
<td>0.14</td>
<td>66</td>
<td>0.44</td>
<td>0.21</td>
</tr>
<tr>
<td>Team Alpha</td>
<td>No Revisions</td>
<td>0.27</td>
<td>0.11</td>
<td>26</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Team Alpha</td>
<td>Revisions</td>
<td>0.37</td>
<td>0.13</td>
<td>48</td>
<td>0.84</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Note: Alpha denotes the estimated response intensities for each individual; Partner’s Alpha denotes the engagement index of the individual’s partner; Team Alpha denotes the team-level index. Hedges’ g used the correction factor described by Hedges (1981) and $r$ denotes the point-biserial correlation. See the main paper for additional details.

References


