JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

*Journal of Educational Measurement*
*Spring 2017, Vol. 54, No. 1, pp. 70–84*

# Measuring Student Engagement During Collaboration

**Peter F. Halpin**
*New York University Steinhardt School*
**Alina A. von Davier**
*ACT, Inc.*
**Jiangang Hao and Lei Liu**
*Educational Testing Service*

*This article addresses performance assessments that involve collaboration among students. We apply the Hawkes process to infer whether the actions of one student are associated with increased probability of further actions by his/her partner(s) in the near future. This leads to an intuitive notion of engagement among collaborators, and we consider a model-based index that can be used to quantify this notion. The approach is illustrated using a simulation-based task designed for science education, in which pairs of collaborators interact using online chat. We also consider the empirical relationship between chat engagement and task performance, finding that less engaged collaborators were less likely to revise their responses after being given an opportunity to share their work with their partner.*

Traditional educational tests target a relatively narrow set of constructs compared to the range of competencies required for student success. One way to address this issue is through the use of performance-based assessments that have better fidelity to the situations in which students learn and are expected to demonstrate their knowledge (e.g., Davey et al., 2015; Pellegrino, Chudowsky, & Glaser, 2001). In this article, we focus on one particular performance context—collaborative problem solving (CPS)—as an avenue for broadening what we can infer about student ability. The use of collaboration and group work for assessment purposes has a relatively long history (e.g., Webb, 1995), and it also features prominently in current initiatives concerning the measurement of "21st century skills" (e.g., Griffin & Care, 2015; Griffin, McGaw, & Care, 2012; National Research Council, 2011; Organisation for Economic Co-operation and Development, 2013; Pellegrino & Hilton, 2012; Stecher & Hamilton, 2014).

An important ingredient that has been added by the recent focus on 21st century skills is the role of information technology in the implementation of more complex assessments, and in the recording of fine-grained data about students' task-related activities. A major challenge in making use of these new data sources is that they often have unclear interpretations with respect to educational and psychological constructs (e.g., National Academy of Education, 2013). In this article, we directly address this issue by describing a statistical framework for using temporally structured data to make inferences about how students collaborate. In particular, we describe the use of point processes (e.g., Daley & Vera-Jones, 2003), and specifically the Hawkes process (Hawkes, 1971a, 1971b; Hawkes & Oakes, 1974), to model statistical dependence in the timing of collaborators' actions. We review

 

details concerning specification, estimation, and goodness of fit of the Hawkes process.

Our review of the Hawkes processes follows quite closely to that of Halpin and De Boeck (2013). The contribution of that article was to develop an expectation maximization (EM) algorithm for a relatively general parameterization of the process. The contribution of this article is to consider the relevance of the methodology to assessments involving CPS. First, we decompose the statistical dependence in a CPS task into components within and across students. We then apply this decomposition to the multivariate Hawkes process, showing that all of the dependence across students can be explained in terms of a model parameter called the *response intensity* (Proposition 1). We also show that the response intensity enjoys a straightforward interpretation as the proportion of a student's actions to which a partner is expected to respond. These results motivate the interpretation of response intensity in terms of engagement among collaborators. We also show how the response intensity parameter can be aggregated to provide a team-level measure of engagement, and outline some preliminary results on its standard error.

Our example addresses online chat between dyads during a simulation-based science game with an embedded assessment (Hao, Liu, von Davier, & Kyllonen, 2015). In this context, the actions to be modeled are chat messages sent between students. Previous work has analyzed student sentiment and strategy during collaboration via the textual content of chat messages (e.g., Howley, Mayfield, & Rosé, 2013; Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015). In contrast, our goal is to infer engagement between students from the temporal characteristics of their chat data, without reference to the content of the messages (for some similar applications, see Barabási, 2005; Crane & Sornette, 2008; Ebel, Mielsch, & Bornholdt, 2002; Halpin & De Boeck, 2013; Oliveira & Vazquez, 2009). We consider the empirical relationships between our proposed measures of engagement and more readily available quantitative summaries of chat data (e.g., the number of words per message), finding that engagement does not correlate highly with these "model-free" indices. We also address the relationship with task performance, finding that less engaged collaborators were less likely to revise their responses after being given an opportunity to share their work with their partner, whereas the model-free indices were not predictive of revisions. We conclude by suggesting avenues for further research. The online Supporting Information contains supplementary materials.

## CPS as a Temporally Complex Task

Davey et al. (2015) discussed task complexity and interconnectedness among task components as necessary characteristics of performance assessments. Of particular relevance to CPS, a distinction has been made between the processes and outcomes of teamwork, with the former denoting a sequence of interrelated actions made by team members (e.g., Brannick, Prince, Prince, & Salas, 1995; von Davier & Halpin, 2013; Webb, 1995). Here we build on these ideas by providing a definition of task complexity that emphasizes the order in which task components are completed. We then show that the overall complexity in a CPS task can be decomposed into parts that depend on individual students, and an additional part that depends on how they

interact with one another. In the following section, this decomposition is used to provide a relatively deep motivation for the engagement indices we propose.

Begin by letting $\mathbf{X}_T = (X_T, X_{T-1}, \ldots, X_1)$ denote a sequence of random variables describing the actions of a student during the completion of a task. For example, in the context of online chat, $X_t$ could be a dichotomous variable indicating whether or not an individual has sent a chat message at time index $t$. Alternatively, $t$ might index the sequence of items on a conventional multiple choice assessment, in which case $X_t$ would denote the item responses and $\mathbf{X}_T$ a $T$-dimensional response pattern. In general, $t$ indexes over some components of a task that are completed in a sequence, and $X_t$ denotes some feature of a student's behavior recorded during the completion of those components. We refer to the $X_t$ as task components, or simply components.

A temporally simple task can be defined as one for which the task components are statistically independent. Obviously, this implies that all tasks obtained by permuting the order of the components have the same joint distribution. The familiar assumption of local independence,

$$p(\mathbf{X}_T \mid \theta) = \prod_{t=1}^{T} p(X_t \mid \theta),$$

can be interpreted as requiring the simplicity of a task, after conditioning on a (possibly vector-valued) latent variable $\theta$. Despite the familiarity of this assumption, we suggest that students' task-related activities during a performance assessment may not be well characterized as conditionally simple. In particular, collaboration is clearly not simple—the order of turns in an interaction cannot be rearranged like the items on a multiple choice test. Thus, rather than pursuing a set of conditioning variables sufficient to ensure local independence (cf. Lord & Novick, 1968, section 16.3), we instead consider alternative modeling strategies that are appropriate for temporally structured data.

To this end, define a complex task as one in which $p(X_t \mid \mathbf{X}_{t-1}) \neq p(X_t)$ for some $t > 1$. This sequential factorization is a common feature of time series methodology, although here we have not required that the indices $t$ denote equidistant time intervals. The factorization implies that a student's actions during a task can depend on his or her previous actions, but not on actions that have not yet occurred. One way of quantifying task complexity is in terms of the Kullback-Leibler (KL) divergence of the joint distribution $p(\mathbf{X}_T) = \prod_{t=1}^{T} p(X_t \mid \mathbf{X}_{t-1})$ from the product of its univariate margins, $q(\mathbf{X}_T) = \prod_{t=1}^{T} p(X_t)$:

$$D[p(\mathbf{X}_T) \,||\, q(\mathbf{X}_T)] \equiv E_p \left[ \ln \frac{p(\mathbf{X}_T)}{q(\mathbf{X}_T)} \right], \tag{1}$$

where $E_p$ denotes expectation with respect to the distribution $p(\mathbf{X}_T)$. KL-divergence is a well-studied quantity (e.g., Cover & Thomas, 2005). In the present context, $D[p(\mathbf{X}_T) \,||\, q(\mathbf{X}_T)] = 0$ describes a simple task. When $D[p(\mathbf{X}_T) \,||\, q(\mathbf{X}_T)] > 0$, some temporal dependence is exhibited by the components, with larger values indicating more dependence.

In application to CPS, we are especially concerned with how the complexity of a task is related to interactions among students. To address this question, let $\mathbf{X} = \{\mathbf{X}_{1\,T_1}, \mathbf{X}_{2\,T_2}, \ldots, \mathbf{X}_{J\,T_J}\}$ denote a collection of $J$ sequences of random variables, with $\mathbf{X}_{j\,T_j} = (X_{j\,T_j}, X_{j\,T_j-1}, \ldots, X_{j\,1})$ denoting the sequence of $T_j$ actions of student $j = 1, \ldots J$ during the completion of a task. Let $q(\mathbf{X}) = \prod_{j=1}^{J} \prod_{t=1}^{T_j} p(X_{jt})$ again denote the assumption that all task components are independent. Then the overall complexity of the task can be decomposed into $J + 1$ parts:

$$
\begin{aligned}
D[p(\mathbf{X}) \,||\, q(\mathbf{X})] &= E_p\left[\ln \frac{p(\mathbf{X})}{q(\mathbf{X})}\right] \\
&= E_p\left[\ln \frac{p(\mathbf{X})}{q(\mathbf{X})} + \ln \prod_{j=1}^{J} \frac{p(\mathbf{X}_{j\,T_j})}{p(\mathbf{X}_{j\,T_j})}\right] \\
&= E_p\left[\ln \frac{p(\mathbf{X})}{\prod_{j=1}^{J} p(\mathbf{X}_{j\,T_j})} + \ln \prod_{j=1}^{J} \frac{p(\mathbf{X}_{j\,T_j})}{q(\mathbf{X}_{j\,T_j})}\right] \\
&= D\left[p(\mathbf{X}) \,||\, \prod_{j=1}^{J} p(\mathbf{X}_{j\,T_j})\right] + \sum_{j=1}^{J} D[p(\mathbf{X}_{j\,T_j}) \,||\, q(\mathbf{X}_{j\,T_j})].
\end{aligned}
\tag{2}
$$

Only the first term depends on the joint distribution of $\mathbf{X}$, and it is equal to zero if and only if the actions of the $J$ students are independent. It can therefore be regarded as describing the portion of overall task complexity that is due to interactions among students, or interindividual complexity. The other $J$ terms depend on the actions of the individual students considered in isolation; these terms reflect intraindividual task complexity.

Equation 2 does not make use of the requirement that the task components are temporally ordered. Next we present an explicit model for temporal dependence and show how its decomposition according to (2) motivates a quantity that can be used to characterize collaboration.

## Modeling CPS With the Hawkes Process

This section informally presents some background on the Hawkes process and addresses its interpretation in the context of collaboration and teamwork. Hawkes (1971a, 1971b) provided the original statement of the process as a conditional intensity function (CIF). The existence and uniqueness of the process were addressed by Hawkes and Oakes (1974) using its so-called branching structure representation, with extensions provided by Brémaud and Massoulié (1996) and Liniger (2009). Maximum likelihood (ML) estimation of parametric models was addressed by Ogata (1978), with more recent work discussing the use of the EM algorithm to facilitate numerical optimization (e.g., Fox, 2015; Halpin, 2013; Halpin & De Boeck, 2013; Lapham, 2014; Mino, 2001; Olson & Carley, 2013; Veen & Schoenberg, 2008; see

Rasmussen, 2012 for a Bayesian approach). A comprehensive treatment of the theory of point processes is provided by Daley and Vera-Jones (2003).

**Review of the Hawkes Process**

There are several equivalent ways of characterizing a point process (see Brillinger, Guttorp, & Schoenberg, 2002; Daley & Vera-Jones, 2003, chap. 3), which we selectively review here to establish notation. Above we mentioned an example in which $\mathbf{X}_T$ was a dichotomous-valued, discrete-time stochastic process with $X_t = 1$ denoting the occurrence of an event at time index $t$. Let $\mathcal{T}$ denote the set of indices $t$ and $\#\mathcal{T}$ its cardinality. An intuitive extension to point processes can be made by assuming that the indices $t \in \mathcal{T}$ are equally spaced over a fixed interval of time $(a, b] \in \mathbb{R}+$ and letting $\#\mathcal{T} \to \infty$. Then $X_t = 1$ denotes the occurrence of an event in the infinitesimal interval $(t, t + dt]$. Another representation is in terms of a random counting measure, $N((a, b])$, which counts the number of events falling in the interval $(a, b]$. Then $N((a, b]) = \sum_{t \in (a,b]} X_t$, and $dN(t) \equiv N((t, t + dt]) = 1$ just in case $X_t = 1$. In what follows, we simplify notation by writing $N((a, b]) = N(a, b]$. We also let $t_i, i \in \mathbb{N}$ denote the sequence of time indices $t$ such that $X_t = 1$. A final consideration is about the cooccurrence of events. The present discussion is limited to orderly point processes, which are defined using Landau's Little-O notation:

$$\text{Prob}(dN(t) > 1) = o(dt).$$

Next we introduce the CIF, which is a convenient mechanism for statistical modeling. Under mild conditions, the CIF uniquely specifies a point process, and it leads directly to important results concerning ML estimation and goodness of fit (see Daley & Vera-Jones, 2003, chap. 7). For an orderly process, the CIF is defined as

$$\lambda(t) \equiv \lim_{dt \to 0} \frac{\text{Prob}(dN(t) = 1 \mid H_t)}{dt}, \tag{3}$$

where $H_t = \{t_i \mid t_i \leq t\}$. Intuitively, $\lambda(t)\, dt$ is an approximation to the Bernoulli probability of an event occurring in the interval $(t, t + dt]$, conditional on all of the events happening before time $t$. In the multivariate case, $\boldsymbol{N}(a, b]$ is vector-valued and each of the $j = 1, \ldots, J$ univariate margins gives the number of a different type of event occurring in the time period $(a, b]$. In application to CPS, we let the univariate margins correspond to the individual students. The CIF then provides a means of describing how the probability of students' actions changes over continuous time as a function of their own previous actions, as well as the previous actions of the other team members.

The Hawkes process is a general framework for modeling the CIF. In their discussion of the process, Daley and Vera-Jones (2003, p. 183) suggest that it comes closest to fulfilling the role of autoregressive models in conventional time-series analyses, with the important restriction that it can only be applied to clustered, or excitatory, processes. We describe this restriction in connection with (6). The CIF of Hawkes process can be written as (cf. Hawkes, 1971a, equation 20).

$$\boldsymbol{\lambda}(t) = \boldsymbol{\mu} + \int_0^t \Phi(t - s)\, d\boldsymbol{N}(s), \tag{4}$$

where $\boldsymbol{\mu} > \mathbf{0}$ is a $J$-dimensional baseline, which can be a function of time but is here treated as a constant, and $\Phi(u)$ is a $J \times J$ matrix of response functions that govern how the process depends on its past. For a single margin, (4) yields

$$\lambda_j(t) = \mu_j + \sum_{k=1}^{J} \int_0^t \phi_{jk}(t-s)\, dN_k(s)$$

$$= \mu_j + \sum_{k=1}^{J} \sum_{t_{ik} < t} \phi_{jk}(t - t_{ik}), \tag{5}$$

where the second line follows by definition of integration with respect to $dN(t)$.

Several restrictions on the $\phi_{jk}(u)$ are required in order for Hawkes process to be well-defined. To begin with

$$\phi_{jk}(u) = 0, u < 0 \quad \text{and} \quad \phi_{jk}(u) \geq 0, u \geq 0. \tag{6}$$

In substantive terms, the first condition requires that an event cannot influence the probability of past events. In physics, this condition is referred to as "physical realizability," and in signal processing it would define $\phi_{jk}(u)$ as a "causal filter." In application to CPS, it means that a student cannot respond to an action that has not yet occurred. The second condition requires that the influence of one event on the probability of another is nonnegative, which leads to the interpretation of the process as excitatory. In application to collaboration, this means that the actions of one student cannot have an inhibitory or suppressive influence on those of any other student. This is clearly a nontrivial restriction. Therefore, it is important to assess the plausibility of this condition for a given data set, prior to application of Hawkes process. We illustrate this with our empirical example.

Restrictions (6) imply that we may write

$$\phi_{jk}(u) = \alpha_{jk}\, f_{jk}(u), \tag{7}$$

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. Further restrictions on the $\phi_{jk}(u)$ may be imposed via various regularity conditions. For example, in the univariate case, the requirement that $E[\boldsymbol{\lambda}(t)]$ is constant yields $\alpha < 1$ (see Hawkes 1971a, equation 9). Liniger (2009, theorem 6.55) provides regularity conditions for the multivariate Hawkes process that require the matrix $A = \{\alpha_{ij}\}$ to have spectral radius (i.e., largest absolute value of its possibly complex eigenvalues) less than one (see also Brémaud & Massoulié, 1996, theorem 7). In the remainder of this section, we focus on providing an intuitive interpretation of the response intensity parameters. The choice of response kernel is discussed in the online Supporting Information.

### Interpretation of Response Intensity

It is apparent from (7) that $\int_0^\infty \phi_{jk}(u)\, du = \alpha_{jk}$. This motivates the interpretation of the response intensity parameter as governing the "overall" or "total" temporal dependence of the actions of student $j$ on those of student $k$. Importantly, this interpretation is independent of the particular form of the response kernel—that is, it is independent of the short-term dynamics of a particular team or task. This is arguably

an important characteristic for any "general purpose" measure of engagement among collaborators.

Another useful expression for $\alpha_{jk}$ can be obtained using the branching structure representation of the Hawkes process, which Hawkes and Oakes (1974) showed to be equivalent to the statement of the model given in (4). In the context of CPS, the branching structure supposes that each action $i$ of student $j$ over the observation period $(0, T]$ can be classified as follows:

(1) The action $i$ of student $j$ is a response to a past action $r$ of student $k$, with $t_{ij} > t_{rk}$. The number of such actions is governed by a Poisson process $N_{jk}(t_{rk}, T]$ with rate $\alpha_{jk} F_{jk}(T - t_{rk})$, in which $F_{jk}$ is the cumulative distribution function (CDF) of the response kernel $f_{jk}$ introduced in (7). Intuitively, each Poisson process corresponds to a term appearing under the double summation in (5), although the formal equivalence is quite difficult to establish (see Hawkes & Oakes, 1974). Note that the student's responsiveness to his or her own past actions is addressed when $j = k$.

(2) The action $i$ of student $j$ is not a response to a past action of any student. The number of such actions is governed by a Poisson process with rate $\mu_j$ given in (5). These events characterize the baseline activity of a specific task or team.

The branching structure representation of the Hawkes process is obtained by requiring that the Poisson processes defined in (1) and (2) above are all mutually independent. For further details, see Rasmussen (2012) or Halpin and De Boeck (2013).

Given the branching structure representation, the expected number of responses made by student $j$ to student $k$ is

$$\bar{n}_{jk} \equiv \sum_{i=1}^{n_k} E[N_{jk}(t_{ik}, T]] = \alpha_{jk} \sum_{i=1}^{n_k} F_{jk}(T - t_{ik}), \tag{8}$$

where $n_k$ is the number of actions of student $k$. This leads to

$$\alpha_{jk} = \frac{\bar{n}_{jk}}{\sum_{i=1}^{n_k} F_{jk}(T - t_{ik})}, \tag{9}$$

and, noting that $F_{jk}(T - t_{ik}) \to 1$ as $T - t_{ik} \to \infty$, we have the intuitive lower bound

$$\alpha_{jk} > \bar{n}_{jk}/n_k. \tag{10}$$

Inequality 10 provides a practical interpretation of $\alpha_{jk}$ as the proportion of actions of student $k$ to which student $j$ would be expected to respond to over a very long interaction. When $j \neq k$, we interpret the response intensity as a measure of student $j$'s engagement with student $k$.

A group-level measure of engagement can be obtained by rescaling the response intensity parameters

$$\alpha \equiv \frac{\sum_{j \neq k} \alpha_{jk} \times \sum_{i=1}^{n_k} F_{jk}(T - t_{ik})}{\sum_{j=1}^{J} n_j} = \frac{\sum_{j \neq k} \bar{n}_{jk}}{n}. \tag{11}$$

This can be interpreted as the proportion of all group members' actions, $n$, to which a response is expected from any other member during a collaboration. In the case of pairwise collaboration, we can use (10) to write

$$\alpha < \frac{\alpha_{12} n_2 + \alpha_{21} n_1}{n_1 + n_2}, \tag{12}$$

which is the team-level engagement index we adopt in our example.

### The Relation Between Response Intensity and Task Complexity

Next we show that the use of response intensity as a measure of engagement can be motivated through a consideration of task complexity as defined in (2). In particular, if no team member is responsive to any other team member (i.e., $\alpha_{jk} = 0$ for all $j \neq k$), then there is no dependence between the actions of the team members. Although the point is quite obvious, we state it as a proposition for explicitness.

**Proposition 1:** *Let* $(0, T]$ *denote the observation period and let* $\mathbf{X}_j = (t_{1j}, t_{2j}, \ldots t_{n_j j})$, $j = 1, \ldots, J$, *denote the event times of the* $j$*th margin of a point process. Assume that the data-generating distribution of* $\mathbf{X} = \{\mathbf{X}_1, \ldots \mathbf{X}_J\}$ *is a Hawkes process as specified in* (4)–(7). *Then* $D[p(\mathbf{X}) \,||\, \prod_j p(\mathbf{X}_j)] = 0$ *if and only if the matrix* $A = \{\alpha_{jk}\}$ *is diagonal.*

The proof is provided in the appendix. A corollary is that the definition of complexity for a CPS task in (2) will include only intraindividual sources of dependence, if and only if $A$ is diagonal. This provides a theoretical motivation for interpreting response intensity as a measure of engagement among collaborators.

### Standard Error of the Response Intensity Parameter

The online Supporting Information provides some details of ML estimation of the Hawkes process, including an informal argument leading to the following lower bound on the asymptotic variance of the ML estimates $\hat{\alpha}_{jk}$ when $j \neq k$:

$$V[\hat{\alpha}_{jk}] > \frac{\alpha_{jk}}{\sum_{i=1}^{n_k} F_{jk}(T - t_{ik})}. \tag{13}$$

It is apparent that the right-hand side is decreasing in $n_k$, which provides some insight into the reliability of the engagement indices proposed in this article. However, as discussed in the empirical example, the lower bound does not compare very well with numerical methods based on the Hessian of the log-likelihood. Future research is needed on topics related to sample size and precision of estimation.

## Empirical Example: The Tetralogue

We now apply the foregoing ideas to an empirical example. Our example was obtained from the Tetralogue (Hao et al., 2015; Liu et al., 2015), which is a simulation-based science game with an embedded assessment. During the simulation, dyads work together via online chat to learn about volcano activity. At various points in the simulation, students are asked to individually submit their responses to an assessment item without discussing the item. Following submission of responses from

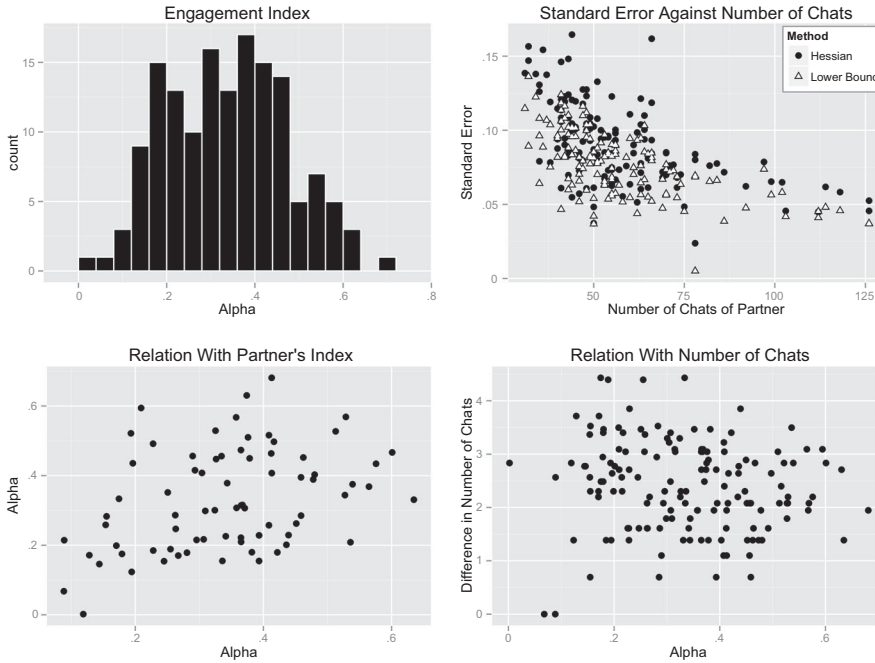


*Figure 1.* Alpha denotes the estimated response intensities from (9). Hessian denotes standard errors obtained via the Hessian of the log-likelihood. Lower bound denotes (13). Difference in number of chats was scaled using the log of the absolute value of the difference.

both students, they are invited to discuss the question and their answers. Last, they are given an opportunity to revise their responses to the item, with the final answers counting toward the team's score.

The data set used for the present analysis contained a total of 268 dyads. Research participants were solicited using the crowdsourcing service Amazon Mechanical Turk and were paired with one another based on their arrival in the queue. The median reported age was 31.5 years, 52.5% reported that they were female, and 79.2% reported that they were White. In addition, 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 results described in this section made use of data from a subset of 74 dyads whose total number of chat messages was in the range $[85, 232]$. The online Supporting Information describes additional details of the overall sample, the screening procedures used to arrive at the analytic sample, and the steps involved in estimating and assessing the goodness of fit of the Hawkes process to the chat data from each dyad.

Figure 1 provides some descriptive information about the estimated engagement indices. In the top left panel, we see that their distribution was approximately symmetrical. The top right panel plots the estimated standard errors against the total

number of the partner's (not dyad's) chats. Using the sample mean of the squared standard errors obtained from the Hessian of the log-likelihood, the marginal reliability was estimated to be .57. The panel also shows the values obtained by plugging-in the sample estimates to the lower bound in (13). While the lower bound follows the overall trend of decreasing with the total number of partner's chats, there were a total of 14 cases in which the lower bound was actually larger than the estimate obtained via the Hessian. It is not clear whether these cases should be counted against the lower bound, or against the use of asymptotic approximation with relatively small sample sizes. As mentioned, further research is needed on these topics.

The lower left panel of Figure 1 displays the relationship between partners' indices ($r = .37$), from which we conclude that team members did not always demonstrate a similar degree of engagement. The lower right panel addresses the relationship between response intensity and one model-free summary: the difference between partners' total number of messages. We also considered several additional model-free summaries of the raw chat data: (a) number of chats sent; (b) total word count of chats sent; (c) average word count per chat message; and (d) differences between partners' total and average word counts. These quantities were log-transformed and absolute values were taken for differences. Perhaps surprisingly, none of these model-free summaries were strongly related to response intensity ($r \in [-.15, .13]$). This provided evidence that our definition of engagement is not simply a "repackaging" of some simpler quantity.

### Relation With Item Revisions

To address whether chat engagement was related to task performance, we considered participants' answers to a total of seven items of the embedded assessment that were available for analysis. As described in the online Supporting Information, the embedded assessment was not suitable for making inferences about learning gains at the level of individuals or dyads. Instead, we focused 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 is an unambiguous aspect of task performance that is plausibly related to chat engagement between partners. However, because some individuals revised one or more responses from "correct" to "incorrect," item revisions did not necessarily correspond to higher or lower scores on the revised responses. It is important to keep in mind this distinction between item revisions and learning gains when interpreting our findings.

We grouped individuals into "Revisions" versus "No Revisions," and considered the mean engagement within each of the two groups. The results are summarized in Figure 2 (see Table 1 of the online Supporting Information for additional details). The most notable finding is that groups in which neither individual made revisions were also characterized by relatively lower levels of chat engagement, as measured by the team-level engagement index. The standardized mean difference was .84, and the 95% confidence intervals on the sample means did not overlap. Despite the shortcomings of the example, this can be taken as reasonable evidence that our measure of chat engagement was related to performance (i.e., item revisions) on the embedded assessment. We also compared the two groups on the various

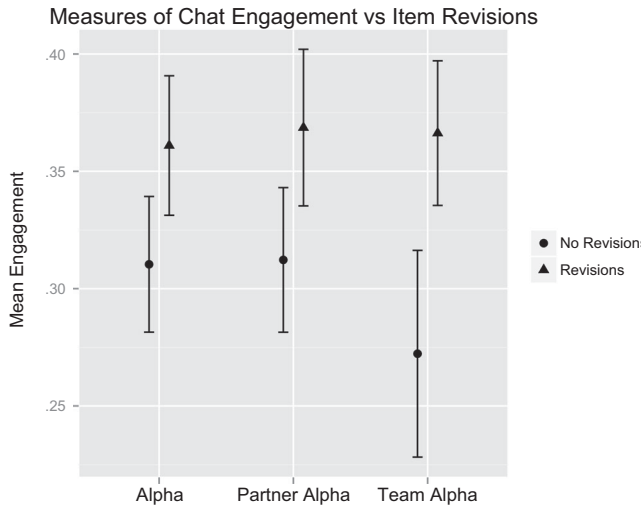


*Figure 2.* Comparison of mean levels of engagement indices for individuals who either did or did not revise at least one response after discussion with their partners. Alpha denotes the estimated response intensities from (9); Partner's Alpha denotes the partner's response intensity; Team Alpha denotes the team-level index in (12). For the latter, the data are reported for dyads, not individuals, and "No Revisions" means that both individuals on the team made no revisions. Error bars are 95% confidence intervals on the means.

model-free indices described in the previous section, finding standardized mean differences in the range $[-.17, .10]$. We conclude that the engagement indices based on the Hawkes process were more strongly related to item revisions than these other summaries.

## Conclusion

This article has addressed the application of the Hawkes process in the context of performance assessments involving collaboration. The model parameters governing response intensity can be used to quantify engagement between specific pairs of team members, as well as an entire team. These measures of engagement have a relatively deep motivation in that they characterize the statistical dependence among the actions of different group members when the data-generating model is correctly specified (see Proposition 1). Moreover, the interpretation of response intensity is quite intuitive; it is the proportion of my partner's actions to which I am expected to respond during a task (Inequality 10). Other indices of chat engagement, such as the number or length of messages, can easily be "gamed" by students, and therefore are not appropriate for assessment applications. However, response intensity seems to come quite close to describing the

kind of behavior we might want students to exhibit as evidence of their ability to collaborate.

Using an empirical example, we examined goodness of fit of the Hawkes process to chat data obtained during a simulation-based science task. The empirical standard errors of the response intensity parameters were examined, as well as the relationships with several other quantities, including a measure of performance (item revisions) on an embedded assessment. Overall, the example supported the conclusion that the Hawkes process is a feasible model for CPS tasks and that the resulting measures of chat engagement are meaningfully related to task performance.

Avenues for improving the application of the Hawkes process to assessments involving collaboration include (a) random effects models for simultaneous estimation over multiple groups; (b) the inclusion of model parameters describing task characteristics, such as the baseline task activity; (c) analytic expressions for standard errors of model parameters; and (d) methods for improving optimization with relatively small numbers of events. Each of these are currently being pursued. Additional lines of inquiry might consider the utility of incorporating text-based features of chat data as time-varying covariates (i.e., the use of marked point processes; see Daley & Vera-Jones, 2003), which could provide a means of integrating the current research with approaches that emphasize the content of chat messages.

**Acknowledgment**

The work was completed when A. A. von Davier was with Educational Testing Service. The opinions presented in this article do not represent the opinions of Educational Testing Service or ACT.

## Appendix

### Proof of Proposition 1

The proof follows simply from writing the log-likelihood of a point process in terms of its CIF (see Daley & Vera-Jones, 2003, chap. 7), and then applying the definition of KL-divergence in (1). The log-likelihood is

$$\ln(p(\mathbf{X})) = \sum_{j=1}^{J} \left( \int_0^T \ln(\lambda_j(s))\, dN_j(s) - \int_0^T \lambda_j(s)\, d(s) \right). \tag{A1}$$

Substituting (5) and (7) into (A1) we have the log-likelihood of the Hawkes process

$$\ln(p(\mathbf{X})) = \sum_{j=1}^{J} \left( \sum_{i=1}^{n_j} \ln \left( \mu_j + \sum_{k=1}^{J} \alpha_{jk} \sum_{t_{rk} < t_{ij}} f_{jk}(t_{ij} - t_{rk}) \right) - \int_0^T \mu_j + \sum_{k=1}^{J} \alpha_{jk} \sum_{t_{rk} < s} f_{jk}(s - t_{rk})\, d(s) \right). \tag{A2}$$

Using the assumption that $A$ is diagonal yields

$$\begin{aligned}\ln(p(\mathbf{X})) &= \sum_{j=1}^{J}\left(\sum_{i=1}^{n_j}\ln\left(\mu_j + \alpha_{jj}\sum_{t_{rj}<t_{ij}} f_{jj}(t_{ij}-t_{rj})\right)\right.\\ &\qquad \left. - \int_0^T \mu_j + \alpha_{jj}\sum_{t_{rj}<s} f_{jj}(s-t_{rj})\,d(s)\right)\\ &= \sum_{j=1}^{J}\ln(p(\mathbf{X}_j)),\end{aligned} \tag{A3}$$

where the second equality follows from the definition of the univariate Hawkes process and implies

$$D\left[p(\mathbf{X}) \mid\mid \prod_{j=1}^{J} p(\mathbf{X}_j)\right] = D\left[\prod_{j=1}^{J} p(\mathbf{X}_j) \mid\mid \prod_{j=1}^{J} p(\mathbf{X}_j)\right] = 0. \tag{A4}$$

Necessity follows since $D[p \mid\mid q] = 0$ requires $p = q$ for any two distributions $p$ and $q$.

## References


Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207–211.

Brannick, M. T., Prince, A., Prince, C., & Salas, E. (1995). The measurement of team process. *Human Factors*, *37*, 641–651.

Brémaud, P., & Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *Annals of Probability*, *24*, 1563–1588.

Brillinger, D. R., Guttorp, P. M., & Schoenberg, F. P. (2002). Point processes, temporal. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of environmetrics* (Vol. 3, pp. 1577–1581). Chichester, England: John Wiley.

Cover, T. M., & Thomas, J. A. (2005). *Elements of information theory*. New York, NY: John Wiley.

Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, *105*, 15649–15653.

Daley, D. J., & Vera-Jones, D. (2003). *An introduction to the theory of point processes: Elementary theory and methods* (2nd ed., Vol. 1). New York, NY: Springer.

Davey, T., Ferrara, S., Holland, P. W., Shavelson, R. J., Webb, N. M., & Wise, L. L. (2015). *Psychometric considerations for the next generation of performance assessment* (Technical Report). Princeton, NJ: Educational Testing Service.

Ebel, H., Mielsch, L.-I., & Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E*, *66*(035103), 1–4.

Fox, E. W. (2015). *Estimation and inference for self-exciting point processes with applications to social networks and earthquake seismology* (PhD dissertation). University of California Los Angeles, Los Angeles, CA.

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills: Methods and approach*. New York, NY: Springer.

Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*. New York, NY: Springer.

Halpin, P. F. (2013). A scalable EM algorithm for Hawkes processes. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology: Presentations from the 77th international meeting of the Psychometric Society* (pp., 403–414). New York, NY: Springer.

Halpin, P. F., & De Boeck, P. (2013). Modelling dyadic interaction with Hawkes processes. *Psychometrika*, *78*, 793–814.

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2015). Assessing collaborative problem solving with simulation-based tasks. In O. Lindwall, P. Häkkine, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the material conditions of learning: Proceedings of the 11th international conference on computer supported collaborative learning* (Vol. 2, pp. 544–547). Gothenberg, Sweden: International Society of the Learning Sciences.

Hawkes, A. G. (1971a). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*, 83–90.

Hawkes, A. G. (1971b). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society, Series B*, *33*(3), 438–443.

Hawkes, A. G., & Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, *11*, 493–503.

Howley, I., Mayfield, E., & Rosé, C. P. (2013). *Linguistic analysis methods for studying small groups*. New York, NY: Taylor & Francis.

Lapham, B. M. (2014). *Hawkes processes and some financial applications* (Master's thesis). University of Cape Town, Cape Town, South Africa.

Liniger, T. (2009). *Multivariate Hawkes processes* (Doctoral dissertation). Swiss Federal Institute of Technology, Zürich, Switzerland.

Liu, L., Hao, J., von Davier, A., Kyllonen, P., & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development* (pp. 344–359). Hershey, PA: IGI Global.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. New York, NY: Wiley.

Mino, H. (2001). Parameter estimation of the intensity process of self-exciting point processes using the EM algorithm. *IEEE Transactions on Instrumentation and Measurement*, *50*, 658–664.

National Academy of Education. (2013). *Adaptive educational technologies: Tools for learning, and for learning about learning* (Technical Report). Washington, DC: Author.

National Research Council. (2011). *Assessing 21st century skills* (Technical Report). Washington DC: Author.

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, *30*, 243–261.

Oliveira, J. G., & Vazquez, A. (2009). Impact of interactions on human dynamics. *Physica A*, *388*, 187–192.

Olson, J. F., & Carley, K. M. (2013). Exact and approximate EM estimation of mutually exciting Hawkes processes. *Statistical Inference for Stochastic Processes*, *16*(1), 63–80.

Organisation for Economic Co-operation and Development. (2013). *PISA 2015 draft collaborative problem solving framework* (Technical Report). Paris, France: OECD Publishing.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know*. Washington DC: National Academies Press.

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

Rasmussen, J. G. (2012). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, *15*, 623–642.

Stecher, B. M., & Hamilton, L. S. (2014). *Measuring hard-to-measure student competencies*. Santa Monica, CA: RAND Corporation.

Veen, A., & Schoenberg, F. P. (2008, jun). Estimation of space-time branching process models in seismology using an EM-type algorithm. *Journal of the American Statistical Association*, *103*(482), 614–624.

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations* (Research Report RR-13-41). Princeton, NJ: Educational Testing Service.

Webb, N. M. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis*, *17*(2), 239–261.

## Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site.

**Online Appendix**

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

**Table 1**. Summary of group differences.

## Authors

PETER F. HALPIN is Assistant Professor of Applied Statistics, New York University Steinhardt School, 246 Greene Street, New York, NY 10003; peter.halpin@nyu.edu. His primary research interests include educational measurement.

ALINA A. VON DAVIER is Vice President of Research and Development, ACT Inc., 500 ACT Drive, Iowa City, IA, 52243-0168; alina.vondavier@act.org. Her main research interest is in developing learning and assessment systems.

JIANGANG HAO is Senior Research Scientist in the Statistical Analysis, Data Analysis, and Psychometric Research Division of Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541; jhao@ets.org. His primary research interests include educational data mining and analytics, virtual performance-based assessment, and assessment of collaborative problem solving.

LEI LIU is Research Scientist at Educational Testing Service, 660 Rosedale Rd, Princeton, NJ 08540; lliu001@ets.org. Her primary research interests include learning and assessment technologies, learning sciences, and educational assessments.