Modeling Player Engagement with Bayesian Hierarchical Models

Robert Sawyer[1], Jonathan Rowe[1], Roger Azevedo[2], James Lester[1]

[1] Department of Computer Science, North Carolina State University

[2] Department of Learning Sciences and Educational Research, University of Central Florida
[1] {rssawyer, jprowe, lester}@ncsu.edu, [2] roger.azevedo@ucf.edu

Abstract

Modeling player engagement is a key challenge in games. However, the gameplay signatures of engaged players can be highly context-sensitive, varying based on where the game is used or what population of players is using it. Traditionally, models of player engagement are investigated in a particular context, and it is unclear how effectively these models generalize to other settings and populations. In this work, we investigate a Bayesian hierarchical linear model for multi-task learning to devise a model of player engagement from a pair of datasets that were gathered in two complementary contexts: a *Classroom Study* with middle school students and a *Laboratory Study* with undergraduate students. Both groups of players used similar versions of CRYSTAL ISLAND, an educational interactive narrative game for science learning. Results indicate that the Bayesian hierarchical model outperforms both pooled and context-specific models in cross-validation measures of predicting player motivation from in-game behaviors, particularly for the smaller *Classroom Study* group. Further, we find that the posterior distributions of model parameters indicate that the coefficient for a measure of gameplay performance significantly differs between groups. Drawing upon their capacity to share information across groups, hierarchical Bayesian methods provide an effective approach for modeling player engagement with data from similar, but different, contexts.

Introduction

Recent years have seen growing interest in player modeling in games. A key challenge in player modeling is devising models of player engagement. Engagement has several core components—these include cognitive, emotional, and behavioral dimensions—as well as complex relationships with related constructs such as motivation and interest (Fredricks, Blumenfeld, and Paris 2004; D'Mello, Dieterle, and Duckworth 2017). By inducing models of player engagement from gameplay data using machine learning, we can develop a better understanding of how players engage and disengage with games (Hadiji et al. 2014; Bertens et al. 2017; Demediuk et al. 2018). Predictive models of player engagement have a broad range of applications, ranging from predicting player churn in online games (Demediuk et al. 2018), understanding how students interact with games for learning (Sabourin and Lester 2014), and driving player-adaptive experience managers for personalizing gameplay (Yu and Riedl 2015).

A key challenge in modeling player engagement is accounting for varying contexts in which player interactions with games occur. The gameplay signatures of engaged players may vary significantly depending on players' traits, where they used the game, when they used the game, and which version of the game they played. There are many open questions regarding how effectively player-analytic models generalize between games, players, and settings. Further, there is limited understanding of how we can leverage data from one context (i.e., one group of players in setting A) in order to improve player analytic models for a different context (i.e., a different group of players in setting B).

In this work, we address this question by modeling player engagement in two considerably different contexts using a multi-task learning framework (Bakker and Heskes 2003). Specifically, we utilize Bayesian hierarchical linear models to predict players' intrinsic motivation from in-game behavior data. Bayesian hierarchical models enable usage of prior distributions for model parameters, which are shared between contexts, as well as posterior distributions for model parameters, which are specific to each context and learned from data. In order to train and evaluate multi-task models of player engagement, we utilize data from player interactions with an educational interactive narrative game for middle school science called CRYSTAL ISLAND. We draw upon two complementary datasets that were gathered from a pair of studies representing different contexts. The first context, which we call the *Laboratory Study*, took place in a controlled laboratory setting with undergraduate students using a baseline version of the CRYSTAL ISLAND game (Taub et al. 2017). The second study, which we call the *Classroom Study*, took place in a middle school science classroom with eighth-grade students using a modified version of CRYSTAL ISLAND that was enhanced to support students' reflection processes. Using this data, we compare a multi-task model of player engagement to both pooled and context-specific models of player engagement induced with the same data. Further, we investigate the uncertainty of the model's parameters, and we compare them across tasks to investigate their predictive value in Laboratory and Classroom Study contexts.

Related Work

Player modeling is critical for understanding how players experience games. A key area of research on player engagement is modeling player churn, which aims to detect player disengagement, or when players stop playing a game, using logs of players' in-game behaviors (Mahlmann et al. 2010; Hadiji et al. 2014). Runge et al. (2014) predicted churn rates in a social game and assessed the business value of churn prediction through a controlled experiment that attempted to maintain players. Xie et al. (2015) expanded upon this work by devising a more generalizable feature representation that used frequency of game events to predict player disengagement. More recently, Demediuk et al. (2018) used mixed effects Cox Regression for survival analysis to predict player churn in League of Legends. Bertens et al. (2017) developed a game churn prediction model from survival ensembles that scales to games with millions of users. Together, this work on data-driven models of player churn has shown significant promise for enriching our understanding of player engagement, but it provides few guarantees about the models' generalizability to new populations of players or new versions of the games. Other work has sought to address this problem by investigating generalized models that use game-independent features, but these methods potentially abstract away information that could be useful within a particular game but are not present in other games (Shaker, Shaker, and Abou-Zleikha 2015).

Multi-task learning and transfer learning provide a family of machine-learning techniques that use knowledge, models, and data from similar tasks to enhance the performance on new tasks. Specifically, multi-task learning tries to learn models for both source and target tasks simultaneously. Transfer learning aims to improve performance on a target task using information from a source task (Pan and Yang 2010). To date, there has been relatively little research on applications of transfer learning in games and player modeling. Snodgrass and Ontanon (2016) used domain transfer to generate game levels for three classic video games. Shaker and Abou-Zleikha (2016) showed that transferring knowledge of player experience between two games is possible through feature replacement between tasks. In our work, there are several features that are shared between multiple contexts, and thus we seek a method that can effectively utilize shared features for each group (i.e., conditioned upon the particular setting, population, and version of the game that players are using).

Figure 1: Screenshot of the infirmary in the CRYSTAL ISLAND educational interactive narrative.

In order to model player engagement across multiple contexts, we utilize hierarchical Bayesian models for multi-task learning. Hierarchical Bayesian models have proven useful across a wide variety of applications, including modeling radon measurements (Gelman 2006), student exam score prediction (Bakker and Heskes 2003), and newspaper sales modeling (Vehtari et al. 2017). In this work, we use hierarchical Bayesian models with data from two versions of an educational interactive narrative game to predict player motivation from logs of in-game actions.

Dataset

CRYSTAL ISLAND is an educational interactive narrative game where players take on the role of a medical field agent who must solve a mystery about an infectious outbreak on a remote island. In the game, players explore a 3D virtual environment, interact with non-player characters, manipulate objects that may have transmitted the disease, conduct experiments in a virtual laboratory, read scientific books and articles, and record their findings in a science notebook. CRYSTAL ISLAND has been used by thousands of middle grade students in K-12 schools in the United States and internationally, and it has been shown to provide significant benefits for science learning (Rowe, Shores, Mott, and Lester, 2011). In this work, we utilize data from two studies involving CRYSTAL ISLAND. Each study involved a different group of students, took place in a different research setting, and centered on a slightly different version of the CRYSTAL ISLAND software.

The *Laboratory Study* was conducted in a laboratory setting at a large mid-Atlantic university with college-aged students ranging from 18 to 26 years old (M = 20.1, SD = 1.6). The original study assigned students to three different experimental conditions, but in this work, we utilize data from only one of the conditions, which involved students using a standard version of the CRYSTAL ISLAND game (Taub et al. 2017). In this study condition, there were a total of 68 students, and after removing 5 due to corrupted or missing game trace logs, the study contained data for 63 students (66.7% female). Participants in this condition played CRYSTAL ISLAND until they solved the mystery (95%), which resulted in a range of total gameplay durations from 35.5 minutes to 160.7 minutes (*M* = 69.5, *SD* = 22.0). During the study, students also completed a pre- and post-test of science content knowledge, where 49 participants (78%) demonstrated positive learning gains with an average normalized learning gain of 0.267 (*SD* = 0.26).

The *Classroom Study* was conducted at a middle school in the mid-Atlantic region within an eighth-grade science class, providing data for 44 students ranging in age from 13 to 14 (*M* = 13.4, *SD* = 0.5). The study was tied to the regular microbiology unit in students' science class, and therefore students received complementary instruction about microbiology prior to exploring CRYSTAL ISLAND. Students played the game until solving the mystery (59%) or two class periods had expired, with a range of total gameplay lasting from 31.4 minutes to 143 minutes (*M* = 79.1, *SD* = 19.9). Students in this study used a slightly modified version of CRYSTAL ISLAND in which reflection prompts were given at specific milestones in the game. These prompts asked students to rate their progress on a scale of 1-10 and reflect on their progress toward solving the mystery. Students took a similar pre- and post-test as the Laboratory Study group, but three questions were removed for being too difficult for middle-school students. Students in the Classroom Study also demonstrated positive learning gains, as 24 participants (55%) achieved positive learning gains with an average normalized learning gain of 0.028 (*SD* = 0.252).

In addition to the science content pre- and post-tests, students in both studies also completed several other attitudinal questionnaires, including the Intrinsic Motivation Inventory (IMI; Ryan 1982). The IMI is a questionnaire that measures participants' subjective experience related to a target activity, and it is grounded in self-determination theory, which is a general theory of human motivation (Ryan 1982). The survey consists of 29 items in which students respond on a 7-point Likert scale, and it has been validated across a range of domains (McAuley 1989). The survey includes seven subscales, but the primary subscale utilized here is the Interest-Enjoyment subscale, which consists of 7 items that provide a self-report measure of students' intrinsic motivation toward CRYSTAL ISLAND. In this work, we seek to devise models of player engagement by predicting student responses on the Interest-Enjoyment subscale of the IMI using predictor features distilled from trace logs of students' in-game behaviors.

To predict player motivation, several features were extracted from the game trace logs generated by CRYSTAL ISLAND. These features consisted of in-game problem-solving behaviors standardized for the duration students spent performing the behaviors relative to their time in the game. Specifically, the features included the following:

- Proportion of time spent conversing with non-player characters
- Proportion of time spent reading books and articles
- Proportion of time spent testing virtual objects
- Proportion of time spent editing in-game diagnosis
- Binary indicator of whether the mystery was solved
- Proportion of gameplay time spent answering reflection prompts (for the Classroom Study group)
- Average response on the in-game reflection prompts (on a 1-10 scale, for the Classroom Study group)
- Final game score, which is a measure created by domain experts to assess students' problem-solving process roughly ranging between +/- 1500 (Rowe et al. 2011)

Each of these was standardized to a zero-mean unit normal distribution. Table 1 presents mean and standard deviations for each of these features prior to standardization for each group.

	Laboratory Mean (Std)	**Classroom Mean (Std)**
Conversation	0.13 (0.027)	0.11 (0.036)
Reading	0.40 (0.083)	0.30 (0.14)
Testing Objects	0.028 (0.016)	0.029 (0.017)
Diagnosis	0.091 (0.040)	0.018 (0.021)
Solved Mystery	0.95	0.59
Prompt Time	N/A	0.083 (0.034)
Prompt Response	N/A	6.59 (1.84)
Game Score	674 (616)	34.4 (760)
IMI Score	4.65 (1.36)	5.25 (1.14)

Table 1. Summary statistics of in-game features and response variable (IMI Score) used for modeling player engagement.

Bayesian Hierarchical Linear Models of Player Engagement

Multi-task learning involves performing multiple parallel tasks (in this case, predicting player motivation in different settings) with a single shared model, taking advantage of structural similarities between the tasks in order to improve generalization. We compare three methods for approaching the multi-task problem: (1) the *Pooled Model*, where all data is pooled into the same group, (2) the *Context-Specific Model*, where each group is fitted with its own regularized model, and (3) the *Bayesian Hierarchical Model*, which aims to fit each group individually while sharing information between groups through shared latent priors. The Pooled Model omits information by treating each group

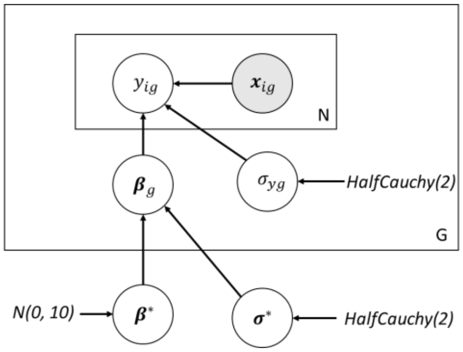

Figure 2. Plate notation of the Bayesian Hierarchical Linear Model with N *total participants and* G *groups.*

equally, potentially underfitting the data. The Context-Specific Model is prone to overfitting each of its respective groups, as each group has its own set of model parameters, increasing the overall model capacity. Thus, the Bayesian Hierarchical Model should provide the best predictive accuracy among these three variations, and it also provides posterior distributions over the model parameters to compare model fit between groups.

We focus on linear models in this work, using in-game behavior features as predictor variables and average IMI interest-enjoyment score as the response variable. Since our datasets are relatively small, linear models provide a natural framework to avoid overfitting, as well as to maintain the interpretability of the model parameters after training. Thus, in the Pooled Model there is one model parameter per feature (plus one intercept) and in the Context-Specific Model and Bayesian Hierarchical Model there are two per feature (one for the Laboratory Study group and one for the Classroom Study group, one intercept for each group).

The pooled and context-specific models use a prior that is normally distributed about zero with a variance of 10. Since linear models with Gaussian priors are used, the Pooled and Context-Specific Models are mathematically equivalent to ridge regression (L2-regularized linear regression) (Murphy 2012). The hierarchical models use a latent variable for the prior mean, which has a hyperprior that follows a normal distribution about zero with variance 10. The latent variable enables sharing of information through penalizing parameters which differ between groups. The variance of the hyperprior over the latent means controls the strength to which the between-group differences in model parameters are penalized; they approach the Pooled Model as the variance goes to zero, and they approach the Context-Specific model as the variance approaches infinity. Setting the prior variance to 10 provides a relatively weak prior and serves to balance the regularization between these two extremes.

All models were trained using Markov chain Monte Carlo sampling in Python with PyMC3 (Salvatier et al. 2016). The plate notation for the Bayesian Hierarchical Model is shown in Figure 2. In this notation, *g* represents the index for a group-specific parameter, * represents a latent variable, *HC* is a Half Cauchy distribution with a scale parameter, *N* is a normal distribution with parameters for mean and variance, and y_{ig} is the *i*th student of group *g*'s interest score while $\boldsymbol{x}_{ig}$ is that student's feature vector composed of in-game actions. Since the Prompt Time and Prompt Response features are not present in the Laboratory data, these features have their own prior (the same as the latent mean vector) instead of the latent variable.

Results

In this section, the predictive accuracy of each model is evaluated under cross validation. The posteriors are used for inference, and the potential for generalization to future tasks is investigated.

Predictive Accuracy

The predictive accuracy for each model was compared using 10-fold cross validation at the player level. For each fold, 1000 MCMC samples were drawn and used to make predictions after omitting the first 500 for burn-in. Burn-in is a common procedure in MCMC sampling that is intended to reduce the potential for sampling from a non-converged Markov chain, and in our case, it comes at little practical cost. The posterior predictive distribution is estimated by making predictions from each of the 1000 sampled model parameters, and the final prediction taken is the mean of these 1000 predictions. This allows the parametric uncertainty of the model to be taken into consideration when making predictions.

Table 2 shows the mean squared error (MSE) of the models averaged over all folds by fold size, as well as cross-validation R^2. Lower MSE indicates a better model, and higher cross-validation R^2 indicates a better model. Cross-validation R^2 is a standardized measure that can be interpreted as the percent decrease in MSE by using predictions from the model rather than a naïve predictor: the mean. It is calculated in the same manner as traditional R^2, except its value can be negative because the residual sum of squares can be greater in magnitude than the total sum of squares, since predictions are made on held-out data. A negative cross-validation R^2 indicates that the predictions from the model were worse than using the mean as the predictor. The MSE for predictions for all students (All), the Laboratory Study group (Lab), and the Classroom Study group (Class) are reported to demonstrate the model's predictive performance within specific contexts.

Table 2 shows that the Bayesian Hierarchical Linear Model outperforms both the Pooled and Context-Specific Models across all students, as well as within each group. A major component of the difference is attributable to the

Pooled Model's weak predictive performance on the Classroom Study group, which suggests that modeling the two groups separately is preferable to take advantage of inherent differences between the two contexts. However, the Context-Specific Model appears to overfit the smaller Classroom Study group, achieving a negative R^2 = -0.251, despite using L2-regularization during training. The R^2 for the Hierarchical Model indicates that its predictions yield a 13.9% reduction (i.e., improvement) in MSE relative to a baseline mean predictor. It similarly yields a 3.4% reduction in MSE relative to the Pooled Model's predictions across all students and a 6.8% reduction relative to the Pooled Model's predictions for the smaller Classroom Study group.

	Pooled	**Context-Specific**	**Hierarchical**
All MSE	1.520	1.677	1.469
Lab MSE	1.583	1.688	1.565
Class MSE	1.430	1.662	1.332
All R^2	0.106	0.0129	0.139
Lab R^2	0.143	0.0826	0.156
Class R^2	-0.0752	-0.251	0.0341

Table 2. 10-fold cross-validation results for predictive accuracy between models and contexts.

Posterior Distributions of Model Parameters

A key benefit of the Bayesian approach to modeling player engagement is that samples from posterior distributions of the model parameters can be used to summarize, compare, and draw inferences about the models. Since the Bayesian Hierarchical Linear Model outperforms the Pooled and Context-Specific models in predictive performance, inferences can be drawn from the induced posterior distributions for the model parameters. Table 3 shows summary statistics for the posteriors for each group in the Hierarchical Model. The row labeled "Uncertainty" lists the estimated standard deviations for the residuals σ_y.

Table 3 reveals that several model parameters differ based upon whether they are associated with the Classroom Study context or Laboratory Study context. For example, the difference in posterior distributions for the Reading Duration coefficient is notable, as this indicates that reading in the Classroom Study group was less predictive of player motivation than in the Laboratory Study group. The uncertainty parameter is also higher in the Laboratory Study group, indicating that model predictions in this context are less certain than in the Classroom Study context, which is further reflected in the overall predictive accuracy of the respective models (Table 2).

The largest difference in parameter estimates between groups is for the Game Score feature. Figure 3 plots the posterior distribution of the Game Score parameter for both the Laboratory Study and Classroom Study groups using 5000 MCMC samples from the full dataset. In 97.6% of these posterior samples, the Game Score parameter for the Laboratory Study group is larger than the Classroom Study group, indicating a statistically significant difference.

	Laboratory		**Classroom**	
	Mean	**Std**	**Mean**	**Std**
Intercept	4.68	0.15	5.22	0.16
Game Score	0.54	0.28	-0.14	0.24
Solved Mystery	0.30	0.23	0.23	0.23
Conversation	-0.28	0.17	-0.15	0.17
Reading	0.04	0.23	-0.18	0.19
Worksheet	0.04	0.21	0.02	0.16
Scanner	0.10	0.23	0.07	0.19
Prompt Time	-	-	0.34	0.19
Prompt Response	-	-	0.40	0.17
Uncertainty	1.19	0.11	1.06	0.13

Table 3. Summary statistics of the posterior distributions of model parameters in the Bayesian Hierarchical Linear Model.

Transfer to Future Tasks

After fitting the Bayesian hierarchical linear model, the latent variables representing the prior mean of each group, which allow sharing of information between contexts, are not used in predictions. However, these latent variables can be used in future transfer tasks, where similar features are available to predict player motivation, as the prior distributions for the model parameters. Using prior means provides a method for addressing the *cold start* problem in player modeling, where a model initially provides poor predictions because it has not yet seen enough data to estimate its model parameters effectively. This issue arises whenever we seek to devise a model of player engagement for a new setting—in the case of Crystal Island, this could be in a museum or home—or with a new population of players, such as high school students. The distribution of the prior means is shown in Figure 4, and it conveys the uncertainty associated with estimating each parameter in a future player modeling task. For example, the distribution of the prior mean for the Game Score parameter has a wider spread than many other parameters, suggesting greater uncertainty about how this feature would transfer to future modeling tasks.

Discussion

The Bayesian hierarchical linear model outperformed both the Pooled Model and Context-Specific Model in predicting player motivation from in-game actions. The Pooled Model performed better in the Laboratory Study setting than in the Classroom Study setting, suggesting that the Pooled Model

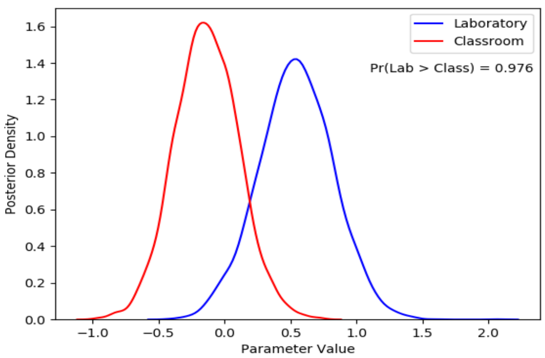

Figure 3. Posterior distributions for coefficients of Game Score for the Classroom Study and Laboratory Study contexts.

was prone to losing key information for modeling player engagement due to differences in how the same features predict player motivation across the two different contexts. For example, Game Score was a strong predictor in both the Laboratory and Classroom studies, but it had a different sign in each context, which the Pooled Model was unable to distinguish between, and thus it evaluated Game Score as a weak predictor. This underscores the value of modeling each context separately, despite the two contexts sharing the same response variable and much of the same feature space.

The Context-Specific Model offers a naïve approach for devising models of player engagement that account for contextual differences by separating the data into separate groups. This approach reduces the effective amount of data available for modeling player engagement in each context, which can lead to overfitting. These downsides were illustrated by the Context-Specific Model's weak performance for the Classroom Study group; the model's predictions were worse than simply using the Classroom Study's group mean. The Bayesian Hierarchical Linear Model regularizes the parameters toward each of the two groups, providing capacity to model each group independently while preventing overfitting from small sample sizes within each group. This enables efficient usage of all available data while simultaneously accounting for contextual differences that may distinguish data from different users, settings, and versions of the game.

We also observed significant differences in the posterior distributions of model parameters between the two groups. The most salient example is the Game Score feature (Figure 3). Since Game Score is a measure of problem-solving performance, the negative coefficient for the Classroom Study group indicates that participants with higher problem-solving performance were predicted to be less motivated toward the game. This negative coefficient could indicate that in the Classroom Study setting, players who solved the mystery efficiently thought that CRYSTAL ISLAND was less interesting than other games or activities they preferred to engage with. The Laboratory Study model did not share the same effect of Game Score on player motivation; a positive coefficient was estimated for Game Score. The developmental differences between the two group populations likely explains the between-group discrepancy in observed Game Scores, as it appears the undergraduate students in the Laboratory Study exhibited more effective problem-solving strategies, which were reflected in higher average Game Scores.

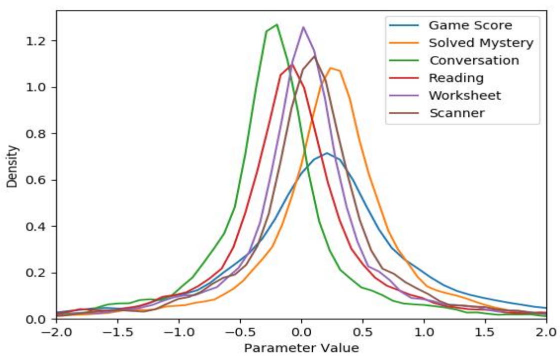

Figure 4. Posterior distributions of the latent variables that are the prior means for the model parameters of each group.

Conclusion

Modeling player engagement across different contexts is an important challenge, as games are increasingly used in different settings by different populations of players, and they are regularly updated with new features and assets. We have presented a model of player engagement that predicts player motivation from in-game actions in two different contexts: a laboratory setting with undergraduate students and a classroom setting with middle-school students. Students in the two studies used slightly different versions of the CRYSTAL ISLAND educational interactive narrative. Results demonstrated that Bayesian hierarchical linear models outperform pooled models and context-specific models in predictive accuracy, and the distribution of model parameters in the Bayesian hierarchical linear model revealed that several features yielded different coefficient estimates across different contexts.

In future work, it will be important to investigate extensions to this framework in order to support run-time models of player engagement and dynamic personalization of player experiences. As more data becomes available, alternative machine learning techniques for modeling player engagement also show promise.

Acknowledgments: This research was supported by funding from the National Science Foundation (NSF) under Grant DRL-1661202 and the Social Sciences and Humanities Research Council of Canada (SSHRC 895-2011-1006). Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or SSHRC.

References

Bakker, B., and Heskes, T. 2003. Task Clustering and Gating for Bayesian Multitask Learning. *Journal of Machine Learning Research* 1(1): 83–99.

Bertens, P., Guitart, A., and Perianez, A. 2017. Games and Big Data: A Scalable Multi-Dimensional Churn Prediction Model. In *Proceedings of the IEEE Conference on Computational Intelligence and Games*, 33–36. Piscataway, NJ.: IEEE.

Demediuk, S., Murrin, A., Bulger, D., Hitchens, M., Drachen, A., Raffe, W., and Tamassia, M. 2018. Player Retention in League of Legends. In *Proceedings of the Australasian Computer Science Week Multiconference,* Article No. 43. New York, NY.: ACM.

D'Mello, S., Dieterle, E. and Duckworth, A. 2017. Advanced, Analytic, Automated (AAA) Measurement of Engagement during Learning. *Educational Psychologist* 52(2): 104–123.

Fredricks, J.A., Blumenfeld, P.C., and Paris, A.H. 2004. School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research* 74(1): 59–109.

Gelman, A. 2006. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics,* 48(3): 432–435.

Hadiji, F., Sifa, R., Drachen, A., Thurau, C., Kersting, K., and Bauckhage, C. 2014. Predicting Player Churn in the Wild. In *Proceedings of the IEEE Conference on Computational Intelligence and Games*. Piscataway, NJ.: IEEE.

Mahlmann, T., Drachen, A., Togelius, J., Canossa, A., and Yannakakis, G. 2010. Predicting Player Behavior in Tomb Raider: Underworld. In *Proceedings of the IEEE Conference on Computational Intelligence and Games*, 178–185. Piscataway, NJ.: IEEE.

McAuley, E., Duncan, T., and Tammen, V. 1989. Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis. *Research Quarterly for Exercise and Sport* 60(1): 48–58.

Murphy, K. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA.: The MIT Press.

Pan, S., and Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10): 1345–1359.

Rowe, J., Shores, L., Mott, B., and Lester, J. 2011. Integrating Learning, Problem Solving, and Engagement in Narrative-Centered Learning Environments. *International Journal of Artificial Intelligence in Education 21*(1-2): 115–133.

Runge, J., Gao, P., Garcin, F., and Faltings, B. 2014. Churn Prediction for High-Value Players in Casual Social Games. In *Proceedings of the IEEE Conference on Computational Intelligence and Games.* Piscataway, NJ.: IEEE.

Ryan, R. 1982. Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory. *Journal of Personality and Social Psychology* 43(3): 450–461.

Sabourin, J., and Lester, J. 2014. Affect and Engagement in Game-Based Learning Environments. *IEEE Transactions on Affective Computing* 5(1): 45–56.

Salvatier, J., Wiecki, T., and Fonnesbeck, C. 2015. Probabilistic Programming in Python using PyMC. arXiv eprint arXiv: 1507.08050.

Shaker, N., and Abou-Zleikha, M. 2016. Transfer Learning for Cross-Game Prediction of Player Experience. In *Proceedings of the IEEE Conference on Computational Intelligence and Games.* Piscataway, NJ.: IEEE.

Shaker, N., Shaker, M., and Abou-Zleikha, M. 2015. Towards Generic Models of Player Experience. In *Proceedings of the Eleventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment,* 191–197. Palo Alto, CA.: AAAI Press.

Snodgrass, S., and Ontanon, S. 2016. An Approach to Domain Transfer in Procedural Content Generation of Two-Dimensional Videogame Levels. In *Proceedings of the Twelfth AAAI Conference on Artificial Intelligence and Interactive Digitial Entertainment*, 79–85. Palo Alto, CA.: AAAI Press.

Taub, M., Mudrick, N., Azevedo, R., Millar, G., Rowe, J., and Lester, J. 2017. Using Multi-Channel Data with Multi-Level Modeling to Assess In-Game Performance during Gameplay with Crystal Island. *Computers in Human Behavior* 76: 641–655.

Vehtari, A., Gelman, A., and Gabry, J. 2016. Practical Bayesian Model Evaluation using Leave-One-Out Cross-Validation and WAIC. *Statistics and Computing 27*(5): 1–20.

Xie, H., Devlin, S., Kudenko, D., and Cowling, P. 2015. Predicting Player Disengagement and First Purchase with Event-Frequency Based Data Representation. In *Proceedings of the IEEE Conference on Computational Intelligence and Games*, 230–237, Piscataway, NJ.: IEEE.

Yu, H., and Riedl, M. O. 2015. Optimizing Players' Expected Enjoyment in Interactive Stories. In *Proceedings of the Eleventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 100–106. Palo Alto, CA.: AAAI Press.