IADIS International Journal on Computer Science and Information Systems
Vol. 14, No. 1, pp. 53-69
ISSN: 1646-3692

RECOMMENDING THERAPEUTIC GAMES TARGETED TO THE INDIVIDUAL NEEDS OF ADULTS WITH AUTISM SPECTRUM DISORDER

Joseph Bills and Yiu-Kai Ng
Computer Science Department, Brigham Young University, Provo, Utah 84602, USA

ABSTRACT

Video games could have potential therapeutic value for individuals on the autism spectrum, but little research has been done on targeting games to the diverse individual needs of adults with autism, and the problem is complicated by the inaccessibility of patient profiles. It is also important to incorporate fun as well as therapeutic value into recommendations. Fun can be estimated by comparing a user's profile of preferred games to the proposed therapeutic games using information from online resources like VideoGameGeek and Wikipedia, even though sorting by therapeutic value is still non-trivial. This can be done by labeling therapeutic games with discrete categories according to their therapeutic value, and sorting games primarily by therapeutic category, and secondarily by estimated fun value. In this paper, we present an approach of using the patient's profile of preferred games as a proxy for their clinical profile, and making game recommendation based on a hypothetical model and updates in response to feedback. This feedback is measured using an ad-hoc questionnaire, which is evaluated on a set of adults with autism spectrum disorder. This model both enables personalized game recommendation from a cold start and allows the learned information to be generalized to other patients.

KEYWORDS

Game Recommendation, Autism Spectrum Disorder (ASD), Adults

1. INTRODUCTION

Autism is a disorder that is defined by impairment in social communication and stereotyped behavior (Bartolome, 2013), which is known to be a spectrum disorder, and those having this disorder have a diverse range of strengths and weaknesses in these areas. For example, deficits in cognitive empathy were once considered to be a universal characteristic of autism, but later research showed this was actually modulated by alexithymia (Bird et al., 2010), which is present

in around half of the people with autism (Hill et al., 2004). Games would be most effective if targeted to the needs of the individual with autism, allowing development of the pivotal skills in which they have a relative deficiency such as social initiation. In practice, therapeutic games target specific areas, e.g., Mindlight targets anxiety (Wijnhoven et al., 2015), and these areas vary in patients with autism (White et al., 2009). Practically, autism therapies work best if tailored to the needs of the individual.

Researchers (Ng & Pera, 2018) have hypothesized that video games could be used as therapeutic tools for people on the autism spectrum. In particular, video games could be integrated into Pivotal Response Treatment (Hiniker, 2013), where essential skills are taught in a naturally motivated manner that results in increased functioning in a wide range of areas (Simpson, 2005). Since many people on the autism spectrum demonstrate strong interest in games (Mazurek et al., 2015), and games by design require mastery of certain skills in order to complete the game process, they represent a natural area to investigate for improving a wide range of skills of people on the spectrum. Most research on the subject has been done on children (Hiniker et al., 2013; Wijnhoven et al., 2015), but there is potential for similar therapy to be applied to adults as autism is a lifelong condition, with Cognitive Enhancement Therapy (CET) proving to have satisfactory effects (Eack et al., 2013). Numerous sorts of skills can potentially be developed, ranging from cognitive to emotional and motor to social, and all of these are important. Development in any of these areas can improve quality of life and productivity.

One problem with implementing a targeted approach is that the medical profiles of autistic patients are confidential, so they cannot be easily accessible. Instead, indirect measures need to be used to construct the patient's profile of strengths and weaknesses. We also want to ensure that games developed for autistic adults are fun for the individual so that they remain engaging. Autism is not rare and there is a high demand for effective autism therapies, so creating new therapies that are both effective and desired by the patients is of utmost importance.

Our solution to this problem is to use a profile of games the patient is interested in as a proxy for the clinical profile by assuming that some sort of underlying correlation exists between the game profile and some game feature that we can measure. An initial hypothetical model about what correlations may exist between games a patient likes and areas of a patient's weakness is used to make recommendations until empirical data has been gathered about which sorts of games are most effective for people with certain profiles. When empirical data is gathered, predictions can be made by matching new user profiles with similar profiles of former users and the games that proved effective for the latter. Our study is the first study to make personalized game recommendations for autistic adults that target both fun and effectiveness using empirical data gathered over the course of the study.

2. RELATED WORK

Granic et al. (2014), who studied positive effects of video games, have shown that commercial games can have positive effects on social skills of their players, both in the short and long term, if the games contain cooperative elements. Their research, however, do not cover what effects may be specific to people with autism.

Eack et al. (2013) have demonstrated that CET can significantly increase cognitive performance in certain areas for adults with autism. The therapy was originally created for adults with schizophrenia, who suffer from similar social skills deficits as adults with autism. The

therapy involves computer-based brain-training exercises, demonstrating the potential for using digital interfaces in the improvement of cognitive abilities. The authors, however, have not investigated if games specifically may be effective.

Mazurek et al. (2015) investigate what autistic adults' opinions on video games are, in terms of their positive and negative effects. One interesting finding is that contrary to popular perception, people have reported more positive social effects than negative social effects. The researchers have also noted the different qualitative elements that influence if someone likes a game. While the study does look into both factors for enjoyment and therapeutic value, they do not come up with a personalized recommendation system like this study aims to do.

Ng and Pera (2018) have developed a system for recommending therapeutic games to adults on the autism spectrum based on personal preference. One of the significant differences of our works and Ng's is that ours accounts for individual differences in what games may be most effective as well as enjoyable, but not the latter.

Yerys et al. (2019) studied how the game Project Evo improves the symptoms of ADHD in children with autism disorder. This is significant because medication is typically ineffective in treating symptoms of ADHD when it is comorbid with autism. This information is relevant as it focuses on using a game to help with specific deficits in patients with autism. The study did not include adults though, nor give information about a recommendation model.

Krach et al. (2018) showed that intervention with the game "The Social Express" caused statistically significant improvement in scores on the Performance Screening version of the Social Skills Improvement System, though the effect size was small for the impacted population. While this study did not mention if any of the participants were on the autism spectrum, the site for the game included many testimonials from parents of children with autism. This game is also only targeted towards children, not adults.

Eichenbaum et al. (2014) showed experimentally how video games could improve cognitive skills in adults, primarily focusing on visual perception and attention. While they did not study adults with autism, they did demonstrate how video games could improve cognitive skills in people with dyslexia. As sensation is often affected in autism, improved performance on visual tasks may be beneficial.

3. OUR PROPOSED MODEL

A game possesses attributes that can be categorized as either being *qualitative* or *therapeutic*, which determine if a game fits a user's taste and will help with his clinical needs, respectively. For example, genre can act as qualitative attribute because some players prefer strategy games while others prefer action games, while the inclusion of brain-training tasks can be a therapeutic trait. These categories are not strictly separate because elements that affect how someone enjoy a game may also relate to potential therapeutic areas. These attributes can be accessed by learning from *labels* at available VideoGameGeek, a social video game website, and other websites that provide structured data on the pages of individual games, or by extracting *phrases* from the descriptions of games on sites such as Wikipedia.

Ideally, qualitative traits and therapeutic traits would be isolated from each other so that a person could be recommended games that match the qualitative traits of the games they already enjoy but contain therapeutic traits that are not currently represented in the games they enjoy. Unfortunately, the two are not strictly different, so our model operates by treating all labels on

a game as defining qualitative traits, and explicitly listing some of these traits as being therapeutic traits. Table 1 gives our hypothetical model. The first column enumerates areas of weakness based on those included in "Bridges We Build: The Art of Making Friends" (Scenicview Academy, 2016) which act as psychological constructs. This set of weakness areas was then restricted to those for which existing games could potentially aid with in isolation after categories with too much overlap to be distinguishable were merged. The next two columns are *labels* and *key phrases* that are hypothesized to be correlated with games that possess attributes that could challenge that area. The last column lists examples of games which have those labels on VideoGameGeek and have the intended attributes.

Table 1. The hypothetical model

Weakness Areas	**VideoGameGeek Label**	**Key Phrases in Gameplay Sections of Wikipedia Articles**	**Example Games**
Communication	Cooperative (Mode)	Team, cooperative	Secret of Mana, Portal 2, Diablo II
Maintaining eye contact	First Person Shooter (Genre)	FPS, first person, first-person, character's view	Halo, Call of Duty, Golden Eye
Responding to others	Hotseat (Mode)	Turn-based, social	Civilization, Advance Wars, Worms
Introducing self/ Making friends	MMO (Genre), Massively Multiplayer (Mode)	Online multiplayer, MMORPG, friends	World of Warcraft, Runescape
Awareness about sensitive subjects/ Social etiquettes	(None)	Etiquette, manners	The Social Express
Handling feedback	Sandbox (Genre), Multiplayer (Mode)	Share, comment, team, creative	Minecraft, Teraria, LEGO Worlds
Resolving conflict	RPG (Genre), Simulation (Genre), Moral Choices (Theme)	Diplomacy, conflict resolution, non-violent	Undertale, Fallout 3
Paying attention	Educational (genre), puzzle (genre)	Focus, details	Brain Age Concentration Training
Difficulty in motor skills (Fine and Gross)	Action (Genre), Wii (Platform), Kinect (Franchise)	Motion controls, typing skills, high difficulty (-turn based), precise timing, requires precision	Mario Teaches Typing (Fine), Wii Sports Resort (Gross)
Sensory difficulty (Listening, seeing)	Rhythm(Genre), Music (Theme)	Sound, queues, graphics	Electroplankton

The hypothetical model has been revised from the initial hypothesis (Bills & Ng, 2019) to reflect changes made after manually reviewing games that were returned according to the previous criteria in the initial hypothesis. The therapeutic traits associated with the labels and phrases that are used to evaluate if the games returned actually contain these traits are included in Appendix 1 (https://bit.ly/2XX5sKf) and notes on which games actually contain these traits are included in Appendix 2 (https://bit.ly/2XOVD1K). This information was used to settle on the set of phrases and labels that are used.

3.1 Sorting Games Using Labels

Since we do not have any information about a patient's clinical profile, we must make inferences about it from the gaming profile that (s)he provides, which consists of a set of games that (s)he enjoys, and the set of labels and phrases found to be associated with those games extracted from VideoGameGeek and Wikipedia, respectively. The hypothesis we are operating from is that if an adult with autism is already playing games that challenge a particular area of weakness, then that area is not a personal weakness for him, and thus it would not be fruitful to recommend games that train only that area. Games will be filtered from a candidate pool of therapeutic games specifically designed to target defined areas of weakness, so that only games that target at least one of the areas that the patient is assumed to have a weakness in are included. It is assumed that a user has a weakness in all the areas to begin with unless their profile matches one of the areas in the model, in which case the user profile is said to have *hit* that area.

In order to diversify results for which games are recommended across patients, areas of weakness will be weighted by how *frequently* they are filtered across all patients, so that an area of weakness will be ranked *higher* if the area occurs *less frequently* among other areas of weakness after being filtered out. This ranking of areas of weakness based on them occurring for less users is equivalent to the area having more users with hits for that area. To calculate exactly how high to rank an area of weakness based on the total number of users who had a hit, the total number of users who had hits must be calculated across all the users. After the games are filtered and categorically ranked by areas of weakness, they are sorted secondarily by expected fun, which is estimated by similarity to game's in the users profile. It is important to note that measuring similarity requires a game to have labels and a description, which must be provided if a therapeutic game is not on VideoGameGeek. The order elements are sorted is captured with RankScore.

$$\textit{RankScore}(\textit{Profile}, \textit{Game}) = \Sigma_{\text{User}}(\text{Hit}_{\text{User, Area-of-Weakness}(\textit{Game})}) + |\text{Labels}(\textit{Profile}) \cap \text{Labels}(\textit{Game})| \,/\, |\text{Labels}(\textit{Profile}) \cup \text{Labels}(\textit{Game})|$$

where $\text{Hit}_{\text{User, Area-of-Weakness(Game)}} = 1$, if that user had a hit in that area of weakness targeted by that game, and 0 otherwise, and Area-of-Weakness(Game) is the area of weakness a particular therapeutic game targets.

Recommending games in order of descending RankScore is equivalent to sorting first by area of weakness, and then sorting the games in each area of weakness by Jaccard similarity. Since areas of weakness are sorted in the same way for each user as that component of RankScore is independent from their profile, computation can be simplified by first sorting a list of games for all users. An algorithm which creates such a preliminary sorted list is listed below, and an example, which shows the sample output from this algorithm based on the given input, is presented in Appendix 3 (https://bit.ly/30CqWtd).

Algorithm. Get_Preliminary_Sorted_List_For_All_Users
Input. A list of users, a list of areas of weakness, and a candidate list of therapeutic games
Output. A list of therapeutic games sorted by area of weakness

1. For each user
   a. Request a list of VideoGameGeek games, *L*, that the user likes as specified in his profile
   b. Initialize a list of *Hits* for each area of weakness, *W*, as **False**
   c. For each area of weakness, *W*
      For each game in *L*
      i. Look up its page, *P*, on VideoGameGeek
         For each label, *A*, in the hypothetical model for that area of weakness *W*
            If *A* is found on *P*, then
               Record the *Hit* for that area of weakness *W* as **True**
      ii. Look up its article, *T*, on Wikipedia
         For each key phrase, *H*, in the hypothetical model for that area of weakness *W*
            If *H* is found in the gameplay section of *T*, then
               Record the *Hit* for that area of weakness *W* as **True**
2. For each area of weakness, *W*
   Initialize Count(*W*) := 0
   For each user, *U*
      If there is a Hit for *U* in *W*, then
         Count[*W*] := Count[*W*] + 1
3. Sort the list of therapeutic games by *decreasing order* based on their counts for the area of weakness it targets so that games in the same of area of weakness remain contiguous, but games within an area of weakness remain unsorted

3.2 Sorting Games Using Text Descriptions

The way that games are sorted by fun is arbitrary, as previous research has been done in this area. In the formula for rank score that was presented in Section 3.1, Jaccard similarity was used due to ease of calculation, but any similarity measure can be used that gives a score between 0 and 1, with 0 being minimally similar and 1 being maximally similar. Another measure that fits these constraints on the range of similarity values is cosine similarity, which requires converting documents to vectors. For this measure, we use the descriptions of the games on Wikipedia as the documents, and extract keywords from these descriptions to form the basis of the vector space. With the extracted keywords, the vector space has a dimension for each keyword that was extracted, and the corresponding vector for a given document has the value of a given component being the number of instances of the corresponding keyword in the document.

To minimize the number of extracted keywords so that the process of computing the similarity measure can be speeded up, we have sorted words in the documents by information gain (Quinlan 1986) and select the top words. This feature selection process finds keywords that correspond closely with designed classes. Words were tokenized solely be separating on the space character, and classes were determined by the *genre* tag on VideoGameGeek. For this, we used the first 6,200 games on VideoGameGeek when sorted by Video Game Rank, a default approach VideoGameGeek sorting the games on their website, which was then filtered down to those for which our scraper could find the corresponding article on Wikipedia and extract its

gameplay description. Many games where given multiple genre tags, in which case the weight of the game was divided equally among each class. For example, if a game were tagged as RPG and Strategy, one half of the weight would be added to the total count of RPG games that is used for the information gain calculation, and the other half would be added to the total count of Strategy games. The issue with this approach is that only the first 30 words were used across multiple games, as then the words unique to the gameplay description of "Pokémon Go" were selected. This was because Pokémon Go was one of two games in our set that was labeled as Augmented Reality and the other, "The Eye of Judgment", was also labeled as Strategy so it only had half weight, giving Pokémon Go majority weight for the class. This resulted in identifying a game as being Pokémon Go as being roughly equivalent to identifying a game as being Augmented Reality in the document, leading words that can be used to identify if a document is the description for Pokémon Go to have high info-gain. The smoothing method for calculating entropy that we used resulted in exact class discriminations being given excessive weight. In addition to these keywords being too specific to generalize to unknown documents, they are redundant as they all represent the same decision for determining whether a document is Augmented Reality, so having one such keyword is as useful as having 70. Having redundant keywords is not only not helpful, but is also harmful for the feature selection as it would result in keywords corresponding with that decision having more weight than the other keywords when cosine similarity is used.

To solve the problem of having redundant keywords that represent the same documents, we improved the feature selection process by organizing candidate keywords into the same data structure as would be used in a decision tree, where each node contains the word with highest information gain for the subset of document set prior to the corresponding decision. After that decision is made, which corresponds with moving to the corresponding child node in the data structure, the document set would be reduced to the documents that either do or do not contain that word. This is not only used for narrowing the set towards document if the tree were being used for classification, but also changes the set that the next node uses to select its word that corresponds with the next decision. Reducing the set of documents after each decision ensures additional keywords corresponding to the same decision will not be selected in subsequent nodes. This results in only one word being selected for each choice, ensuring that redundant terms are removed. To use the tree structure to extract useful keywords, the tree is searched *breadth first* as words that were higher in the tree are less specific than those lower down, and the word whose presence or absence is decided upon at each node is added as a keyword. To further improve on feature selection, *tokenization* is done by removing non-alphabetic characters, converting words to lower case, and adding their Porter stems to the set of words that represent a document. This is important as several of the keywords are just variations on each other, and equating keywords may result in them being detected in more documents. This resulted in less of the keywords being overly specific as the tokenized words in general were less likely to only occur in a single document. The 100 keywords generated from this method are also in Appendix 4 (https://bit.ly/30yv8dw).

3.3 Empirical Data

As it stands, there is no known research evaluating the relationship between arbitrary video game categories such as the labels on VideoGameGeek and these areas of weakness, so the proposed relationship is strictly hypothetical. However, this hypothesis can be tested by

gathering empirical data. *Empirical data* is collected by measuring how users show improvement in these areas based on responses to questions in a questionnaire. An initial baseline for a patient can be established by having them answer an ad-hoc questionnaire that uses questions gleaned from psychometrically evaluated tools, such as the Autism Quotient (Stevenson & Hart, 2017) or similar custom questions that have been labeled according to the areas of weakness we defined. Questions are most useful if they have been confirmed to be reliable, so that changes reflect real change rather than just random variation. Examples of such questions are found in Table 2 (and for the complete set of questions, see Appendix 5 https://bit.ly/2G8wd4b), with all questions either taken from the Autism Quotient or originally devised. For each question, the patient is asked to rate themselves as "*Strongly Agree*", "*Agree*", "*Slightly Agree*", "*Neither Agree Nor Disagree*", "*Slightly Disagree*", "*Disagree*", and "*Strongly Disagree*". For each response, a value between -3 and 3 will be assigned depending on whether the question positively or negatively correlates to that area of weakness. The total score for an area of weakness is the sum of the score for each individual question relating to the area. Note that because these tools were evaluated using our own constructs this can only be used as a baseline, not as a substitute for a clinical profile. Without external validity only relative improvement can be measured, not absolute scores, but relative improvement is enough to validate our model.

Table 2. The sample questions in questionnaire used by the proposed model

Weakness	**Questions**	**Correlation**
Communication	I frequently find that I do not know how to keep a conversation going.	+
	In a social group, I can easily keep track of several different people's conversations.	-
	When I talk on the phone, I am not sure when it is my turn to speak.	+
	I am good at social chitchat.	-
	I find it difficult to work out people's intentions.	+

Patients can be reassessed using the same questionnaire after each recommended therapeutic game they have played in order to monitor how their responses have changed. Based on their responses, a score can be recalculated for each area of weakness, and performance is defined as the updated score minus the baseline score. This information can be used to validate the hypothesis or update the model. If improvement was found in a specific area of weakness, the therapeutic games the patient played can be labeled as having *affinity* with the profile, and the targeted skill should be *filtered out*. This **affinity score** for a user is represented as a vector with dimension for each candidate therapeutic game. If no improvement was found, the game, but not the weakness area, will be labeled as having *negative affinity* with the profile and *filtered* from future recommendation. As the process is iterated, the filtering will narrow, ensuring that all areas of weakness will eventually be considered, and they would be exposed to an effective therapeutic game if any exist. To incorporate fun as well as therapeutic value into the data, the patient will also be asked how much he has enjoyed a game, where they either claim they liked the game (1), disliked the game (-1), or were indifferent (0). Based on this another value will be added to the affinity score, but one with a smaller absolute value than that is assigned based on

whether the therapeutic game was effective, so half the rating is added. Since the exact value for the fun component is irrelevant as it has no impact on order, one-half is chosen for simplicity, and this is done in order to ensure that therapeutic value has more weight.

$$Affinity\_Score(User, Game) = Sgn(\text{Updated\_Score}_{\text{Area of Weakness}(Game),User} - \text{Baseline\_Score}_{\text{Area of Weakness}(Game),User}) + Rating(User) / 2$$

For example, assume that a player has a baseline score of 3 in Communication, and scores 5 in that category after playing a therapeutic game that targets communication. His score improved by two points, which is positive and thus positive affinity is registered, contributing a value of 1 to the affinity score. He also rated that he liked the game (1), which adds 0.5 to the score. The affinity score with the therapeutic game is calculated as being 1.5 overall, and all other therapeutic games targeting communication will be filtered out.

When there are either no more therapeutic games to recommend to a player, or a set period of time (e.g., 2 weeks, long enough for incremental effects to be seen, but short enough that empirical data can be gathered at a reasonable rate) has passed since this user had his first game recommended to him, the player's profile is considered to be *validated*. After enough profiles have been validated to make predications, the hypothetical model can be abandoned for certain users as we will now have empirical data for training a machine learning model that can be used to make more accurate predictions than the hypothetical model by predicting affinity scores for a new user's profile. Whether or not there is *enough* data to make a prediction for a given user is measured by whether the net similarity between the new user's profile and all validated profiles exceeds a certain threshold value, which is defined below. The net similarity is defined as the sum the similarity scores between the new user and each validated profile. Between any two profiles, a *similarity score* is defined in order to gauge if a prediction could be made from existing data. This similarity score is based on the entire set of labels gleaned from VideoGame- Geek for all the VideoGameGeek games that were played, not just those that were included in the hypothetical model, as well as significant monograms and bigrams extracted from the Wikipedia descriptions of those VideoGameGeek games. While arbitrary functions could be used, we calculate the similarity score between two profiles using *Jaccard similarity* as

$$Similarity\_Score(A, B) = \frac{|\,(\text{Labels}(A) \cup \text{Phrases}(A)) \cap (\text{Labels}(B) \cup \text{Phrases}(B))\,|}{|\,\text{Labels}(A) \cup \text{Phrases}(A) \cup \text{Labels}(B) \cup \text{Phrases}(B)\,|}$$

where A and B are any two user's profiles, Labels(X) is the union of the labels across all games in X's profile, and Phrases(X) is the union of the selected phrases across all VideoGameGeek games in X's profile.

If the net similarity of a user profile across all validated profiles is greater than a threshold value, then the game recommendations for the user will be filtered based on calculated affinity to that user rather than by the hypothetical model. The *threshold value* should be dependent on the machine learning model and the similarity measure. As the maximum value for similarity between any two profiles is one and thus the maximum possible sum similarity is just the number of validated profiles, a simple way to estimate the threshold is just use the minimum number of instances to train the model. The minimal number of instances varies depending on what machine learning model is used. Here we decided to use K-Nearest-Neighbors algorithm, which requires at least K+1 instances to decide. We have also considered using a decision tree,

which has no absolute minimum, but generally requires more instances to work effectively, with the rational given in Section 3.4.

$$Threshold\_Value = Maximum\ Possible\ Similarity\_Score \times Minimum\ Number\ of\ Instances\ Needed\ To\ Use\ Model$$

Algorithm. Profile_Validation
Input. A user and a list of sorted, recommended therapeutic games
Output. A vector of Affinity Scores
1. Establish the user's Baseline Score in each Area of Weakness using the questionnaire answered by the user
2. Initialize the user's *Affinity Scores* with each game to *zero*
3. While there are still therapeutic games in the recommendation list and time remaining for the user to be evaluated (< 2 weeks)
   a. Remove the first game, *G*, from the recommendation list and recommend *G* to the User
   b. Wait for the user to complete playing *G*
   c. Prompt the user to rate *G*
   d. If the user likes *G*, then //+- 0.5 based on like, absolute values < 1 so it has less impact
      Set *Affinity_score*[*G*] := 0.5
      Else
      Set *Affinity_Score*[*G*] := -0.5
   e. Have the user take the questionnaire again
   f. Calculate the Relative Improvement in the Area of Weakness *G* targeted and save it as New Score
   g. If improvement was found, i.e., *New Score* > *Baseline Score*, then //+- 1 based on improvement
      i. *Affinity_Score*(*G*) := *Affinity_Score*(*G*) + 1
      ii. Filter out all the therapeutic games that targeted the particular Area of Weakness
      Else
      iii. *Affinity_Score*(*G*) := *Affinity_Score*(*G*) - 1

3.4 Data Clustering

If enough data, i.e., at least 20 instances as there are ten areas of weakness and two extremes for affinity scores, is collected, a decision tree could be trained to predict whether each individual therapeutic game would be useful to a user based on the user's profile. However, with limited data due to the slow empirical process another method other than decision trees may be more efficient. (There would have to be at least as many data instances as there are labels before entropy of a decision tree can be meaningfully interpreted.) Hence, we adopt a semi-supervised model that first clusters validated profiles with similar affinity scores and then associate an unvalidated profile using K-Nearest Neighbors (KNN), which works better with limited information as it assumes users fall into a smaller number of classes. Here KNN uses unweighted voting, and in the event of a tie, the furthest neighbor is excluded.

In our case, KNN with an arbitrary small *k*-value, such as four, on *k*-1 clusters can make meaningful predictions sooner as it assumes simpler boundaries between qualitatively distinct categories of user profiles. For this model, the threshold is *k*+1 to ensure that at least one validated profile will not be included in the set of nearest neighbors. KNN predicts which cluster

a new user profile should be assigned to and returns the vector of affinity scores associated with the cluster. It does this by finding the K-nearest neighbors, defined as the *k* profile's with the highest similarity scores to the new profile. Clusters are determined using a variation on K-Means for which distance is defined as the cosine similarity between the affinity scores of a validated game and an estimated mean. This algorithm works by first randomly assigning each validated profile to a cluster. Then the mean of each cluster is calculated using the mean affinity scores of each associated profile. After the means are calculated, each profile is assigned to the cluster corresponding with the closest mean, and the clusters are recalculated. This process is repeated until no changes occur. Once profiles are assigned to clusters, that cluster can be added as a class that can then be predicted for unvalidated profiles using KNN, and from the cluster, a vector of affinity scores is predicted for the unvalidated profile that was assigned to the cluster's class. The predicted affinity scores returned for a cluster will be the same as the *mean* that defines the cluster. Therapeutic games will be ranked by affinity score and the highest-ranking game will be recommended.

Example 2. Suppose there are six validated profiles and K-Means separates them into three clusters, each with two games, where those two games have higher cosine similarity between their affinity scores than with any game outside the cluster. When a new user arrives, his unvalidated profile is compared to all of the validated profiles. As an example, assume two of the 4 (= *k*) nearest profiles are in the same cluster (see the square shaped profiles in Figure 1), so that cluster is predicted. The average of those two profile's affinity scores is returned and assigned as the affinity scores of the unvalidated profile so it may be used for sorting.

Figure 1. Visual representation of Example 2

Each user's game profile is represented by position in space. Clusters are enclosed by curves and are given a unique shape. The nearest neighbors point to the new unvalidated profile whose affinity scores are initialized to zero and is represented as black. The arrow coming out of the new profile shows what cluster it will be initially assigned to (i.e., the cluster with squares) and what its affinity scores will be set to before games are sorted and empirical data is gathered.

The complete recommendation system combines the use of the hypothetical model for preliminary sorting and filtering, description-based sorting, empirical validation, thresholding, clustering, and class prediction. First, the preliminary list of games is sorted using the candidate

list of games. After which it is possible to recommend games to each user, which can be done in parallel, but should be staggered so the empirical data from early users can help later, new users. Each user's profile is first compared to the validated profiles to see if there is enough empirical data to make predictions for them. If their net similarity exceeds the threshhold, KNN is used to predict their affinity scores with each candidate therapeutic game, and then these games are sorted by their affinity scores. Otherwise, the games are primarily sorted by the preliminary sort, and then filtered using the hypothetical model and secondarily sorted by estimated fun using description-based sorting. Regardless of how the games were sorted, sorted games are recommended in order unless their area is filtered during the validation process, during which empirical data is also gathered. After the validation process is complete, the profile is added to the list of validated profiles, which are then clustered using K-Means.

Algorithm. Recommend_Games
Input. A list of users, and a list of candidate therapeutic games
Output. A sorted recommendation list for each user
1. Initialize the set of Validated_Profiles as an empty set
2. Preliminarily sort the list of games for all users using the Get_Preliminary_Sorted_List_For_All_Users Algorithm
3. For each user, *U*
   a. Initialize *Net_Similarity* to *zero*
   b. For each Validated_Profile (a set that will increase in size after each user)
      i. Calculate the Similarity score between the Validated_Profile and *U*'s Profile
      ii. Add this *Similarity score* to the *Net_Similarity Score*
   c. If Net_Similarity Score > Threshold, i.e., *Net Similarity* > *K* + 1, then
      i. Use K-Nearest Neighbors to assign the *U*'s Profile to a cluster of validated profiles
      ii. Set the Predicted score for each game in the recommendation list as the average Affinity score in the cluster
      iii. Sort the Recommendation_List by Predicated scores in decreasing order
      Else
      iv. Initialize *U*'s Recommendation_List to the Preliminary Sorted List
      v. For Each Area of Weakness
         (a) If *U* has a *hit* in that area, then
            Filter all games targeting that area out of the User's Recommendation List
      vi. Further sort the List by comparing games that target the same Area of Weakness and rating those with higher similarity scores to the user's profile higher
   d. After the Recommendation_List is filtered and sorted, make recommendations using the Profile_Validation algorithm
   e. Record empirical data by adding *U*'s profile and affinity scores to the list of Validated_Profiles
   f. Use K-Means to Divide the Validated_Profiles into *k*-1 Clusters

Example 3. Assume that there are seven users. For the first user, there is no validated profiles, so the net similarity must be zero. Since zero is less than the threshold, we sort the list of therapeutic games using the hypothetical model. In fact, for the first five users, the set of validated profiles will have less than five profiles in it, so the net similarity score will necessarily be less than five, which is the threshold. As such, they will all receive recommendations based on the hypothetical model by taking the preliminary sorted list and then filtering it based on

their preferences, and finally applying a secondary sort. After each user profile has its list filtered and sorted, the corresponding users are recommended games until that process terminated, then they are added to the set of validated profiles with the affinity scores that were just calculated. With the sixth user, there is now five validated profiles, so it is possible for the net similarity score equal to five, but only in the case where new profile is identical to all the existing profiles so it has the maximum similarity score of one with each individual profile, and it cannot exceed five. Since the threshold still cannot be breached, we know the sixth user's profile is given the same preliminary sorted list as the previous five, and it proceeds the same. After this point, we assume that k-means ends up clustering the validated profiles as described in Example 2. We then assume that for the seventh user, the net similarity score is found to be greater than the threshold, causing KNN to be applied. It is assigned to the cluster described in Example 2, with those predicted affinity scores. The candidate games are then sorted based on these affinity scores, and after they are sorted, recommendations and empirical data are gathered in the same manner. No filtering occurs. This validation process results in a net vector of affinity scores. This new vector is recorded with the validated profiles. Finally, K-Means is applied again, and results in a new clustering, for example, with the previous validated profiles ending up with the same clusters, but the new validated profile ended up being clustered in the circle group rather than the rectangle group as shown in Figure 1 as it was originally assigned. □

4. VALIDATION TESTS

In order to test the validity of the system described in this paper, an empirical study must be conducted. If fact, we can assess some other aspects of the model without gathering empirical data. One way to do this is look at the internal correlation of the hypothetical model, seeing if the games on VideoGameGeek with the labels from the hypothetical model correspond with games containing key phrases on Wikipedia for the same area of weakness. In general, if any two measures are highly correlated, they likely measure the same construct, though they should not be perfectly correlated as in that case using multiple measures is redundant. For each area of weakness that an individual game may help strengthen, a hit can be separately calculated using the labels in VideoGameGeek and key phrases from Wikipedia. Correlation between hits, VideoGameGeek hits, and Wikipedia hits in each area of weakness over a collection of games can be calculated using the Pearson Φ *coefficient*, where each count is based on the number of games that were *hit* or *missed* on either VideoGameGeek or Wikipedia.

$$B_a = \Sigma_g (V_{g,a} \times W_{g,a}),\ V_a = \Sigma_g (V_{g,a} \times (1 - W_{g,a})),\ W_a = \Sigma_g ((1 - V_{g,a}) \times W_{g,a}),$$
$$N_a = \Sigma_g ((1 - V_{g,a}) \times (1 - W_{g,a}))$$
$$\Phi_a = (B_a \times N_a - V_a \times W_a) / \mathrm{Sqrt}((B_a + V_a) \times (B_a + W_a) \times (N_a + V_a) \times (N_a + W_a))$$

where $V_{g,a}$ is the game g matched that area a on VideoGameGeek and $W_{g,a}$ is the game g matched that area a on Wikipedia, and B, V, W, and N stand for total instances of cases where games matched both VideoGameGeek and Wikipedia (B), only VideoGameGeek (V), only Wikipedia (W), or neither (N), respectively. A value greater than 0.5 suggests strong correlation and thus a valid construct, while less than 1 shows that it is not redundant.

4.1 Group Evaluation

The effectiveness of the entire procedure can be evaluated by clinically testing each patient with autism spectrum disorder who volunteer for the project and who is not assigned to testing just the hypothetical model. Each patient here is evaluated at the start of the experiment, and again at the end of a trial period. They are evaluated by a clinician according to some numerical measure that has already been studied so that the degree of change required for statistically significant improvement is already known. To minimize bias, the evaluators must be blind to what group a patient was assigned to. This trial period is around a month to give patients time to work through all of the games. These patients are randomly assigned into three groups: a *control group* that is recommended no games, a *baseline group* that play games that are recommended solely based on personal enjoyment, and a *test group* that have games recommended for them using the system described here. Ideally each group would have at least nine people in it so that clustering can be tested. To ensure that clustering is tested, the patients in the 3rd group should be staggered so one patient starts every three days, ensuring that a validated set is built up before later patients start receiving recommendations. Even though the patients are staggered, the total number of hits is calculated for the entire group at the beginning. Each group can be reassessed to determine what percentage of the people in them show significant improvement, with the clinicians reporting the degree of improvement for each patient according to whatever measure they choose, and then the patient is labeled as having significant improvement or not based on known properties of the distribution for that measure. The resulting percentages can be reported without revealing any personal information by ensuring that only the computer system that calculates the percentage of people who showed significant improvement knows which group people were assigned to, and then only the percentages are reported to the analysis team.

Significant testing in one group over another is done using Fisher's exact test. The frequency values in the following table can be recovered by multiplying the percentage with the sample size.

Frequency Table	#Control Group	#Test Group
# Significant Improvement	*w*	*x*
# No Significant Improvement	*y*	*z*

$$p = \mathrm{C}(w + x, w) \times \mathrm{C}(y + z, y) / \mathrm{C}(n, w + y)$$

where C(*a*, *b*) is combinations of *b* from *a*, and $n = w + x + y + z$

4.2 Significance Test

This test was chosen due to still being usable with small sample sizes. If $p < 0.05$, which is the standard *p*-value in psychology, then the null hypothesis that the two involved groups were equally effective is to be rejected. If the null hypothesis is rejected and the proportion of individuals with significant improvement in the test group is greater than the control group, then the test group is to have significant improvement over the control group.

If no significant improvement can be seen in either the *baseline* or the *test group* over the *control group*, then the preliminary hypothesis that the recommended games are effective therapy must be rejected. In this, no conclusion can be made about the relative advantage of a

recommendation system that targets to individual weaknesses over one that does not. The fundamental construct of what games are used and how they are administered must be revised before it can be determined if this targeted system gives a relative advantage over an untargeted one. If the *test group* significantly outperforms the *baseline group*, then it can be concluded that this system works better for targeting individual weaknesses by *rejecting* the *null hypothesis*.

5. PRELIMINARY RESULTS

For the first test, we computed Pearson Φ coefficient in each area for the first 1000 games on VideoGameGeek (excluding 82 games for which the Wikipedia page could not be found):

Communication: 0.26	Eye_Contact: 0.17	Responding: 0.13	Etiquette: 0.11	Initiation: 0.09
Feedback: 0.23	Conflict: -0.05	Attention: -0.06	Motor: -0.02	Sensory: 0.14

Note that for these calculations, each possible outcome had its count increased by one. This was so etiquette would have a defined value despite not having any labels. This resulted in a slightly higher Pearson Φ coefficient for most of the other variables, but not significantly so. In most cases, the correlation is positive, but small.

For the second test, we recorded total VideoGameGeek hits and Wikipedia hits for each area of weakness, where one hit was registered for each game that contained one of the labels or phrases.

VideoGameGeek:Hits

Communication: 155	Eye_Contact: 92	Responding: 56	Etiquette: 0	Initiation: 16
Feedback: 485	Conflict: 237	Attention: 54	Motor: 334	Sensory: 16

Wikipedia:Hits

Communication: 231	Eye_Contact: 18	Responding: 114	Etiquette: 33	Initiation: 89
Feedback: 255	Conflict: 31	Attention: 155	Motor: 20	Sensory: 108

While the number of hits varied greatly from one category to the next, they all achieved significant representation, so we are not worried about any of the terms being too restrictive to possibly get any results. Some of the labels appear to be overrepresented though, and could be restricted by changing the definition of a hit to be more strict.

We have tested the reliability questionnaire on a control group of 11 patients who were enrolled at ScenicView Academy, a local school for adults with autism. To see how they respond to the questions initially and a month later. While not all patients responded at the ends of the same month interval, all the response intervals were taken between April 19, 2019 and June 2, 2019. For each area of weakness, a total score was calculated by summing the responses to the questions for each area of weakness. These total scores are integers vary from -15 to 15. The differences in these scores was calculated for each user between the first and second month, and then the mean and standard of deviation across these differences. These values are given below.

	Communication	Maintaining Eye Contact	Responding to Others	Introducing Self/Making Friends	Awareness about sensitive subjects/social etiquettes
Mean Difference	-0.545454545	-1.1818181	1.81818181	0.363636363	-1.09090909
Standard of Deviation	2.339386089	4.3317013	1.32801971	1.629277587	2.84445233
	Handling Feedback	Resolving Conflict	Paying Attention	Difficulty in Motor Skills	Sensory Difficulty
Mean Difference	0	1	-0.3636363	-1.18181818	1.636363636
Standard of Deviation	4.335896678	3.06594194	2.0135901	3.81623326	3.880018744

From these statistics, a statistically significant improvement in scores can be defined as difference that is two standards of deviation greater than the mean. This is approximately equivalent to a *p*-value of 0.5 when assuming a null hypothesis at the mean under a normal distribution.

We did have a test group of six patients by comparing their communication scores before and after playing a therapeutic game, but no significant improvement was found in any of the patients. Note that failure to reject the null hypothesis is not acceptance of the null hypothesis, especially with such a small sample size and expected effect size, so no conclusions can be drawn on the effectiveness of this game.

6. CONCLUSION

Previous research on recommendation systems for therapeutic games focused only on personal preference for a game (Ng & Pera, 2018). By recommending games based on individual therapeutic value as well as qualitative value, we can ensure that patients will not only have fun, but that they will also develop the skills they need. We solve this problem by finding a correlation between game preference and areas of weakness. Then recommendations can be made primarily based on that correlation, and secondarily on games that are similar to those that are enjoyed. By targeting games to a patient's individual needs, the games can be used more effectively, leading to a more efficient mastery of essential skills. In addition, we have also recommend games based on validated user profiles for a new user, which enables us to recommend games quickly without using the hypothetical model as previously introduced.

REFERENCES

Bartolome, N. et al, 2013. Autism Spectrum Disorder Children Interaction Skills Measurement Using Computer Games. *Proceedings of the 18th* IEEE *International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games (CGAMES)*, pp. 207-211.

Bird, G. et al, 2010. Empathic Brain Responses in Insula are Modulated by Levels of Alexithymia but not Autism. In *Brain: A Journal of Neurology*, Vol. 133, Issue 5, pp. 1515-1525.

Eack, S. et al, 2013. Cognitive Enhancement Therapy for Adults with Autism Spectrum Disorder: Results of an 18-Month Feasibility Study. In *Journal of Autism and Developmental Disorders*, Vol. 43, No. 12, pp. 2866-2877.

Eichenbaum, A., Bavelier, D., & Green, G. S. (2014). Video Games: Play that Can Do Serious Good. American Journal of Play, 7(1), 50. Granic, I. et al, 2014. The Benefits of Playing Video Games. In *American Psychologist*, Vol. 69, No. 1, pp. 66-78.

Hill, E. et al, 2004. Brief Report: Cognitive Processing of Own Emotions in Individuals with Autistic Spectrum Disorder and in their Relatives. In *Journal of Autism & Developmental Disorders*, Vol. 34, No. 2, pp. 229-235

Hiniker, A. et al, 2013. Go Go Games: Therapeutic Video Games for Children with Autism Spectrum Disorders. *Proceedings of the 12th International Conference on Interaction Design and Children (IDC)*, pp. 463-466.

Joseph Bills and Yiu-Kai Ng, Targeting Therapeutic Games to Adults with Autism Spectrum Disorder. In Proceedings of the 9th International Conference on Internet Technologies & Society (ITS 2019), pp. 19-26, Hong Kong, February 8-10, 2019.

Krach, S. K., Doss, K. M., Highsmith, D., Brown, L. S., & McCreery, M. P. (2018). Can computers teach social skills? examining "The social express". (). Annual Convention, Atlanta, GA.:

Mazurek, M. et al, 2015. Video Games from the Perspective of Adults with Autism Spectrum Disorder. In *Computers in Human behavior*, Vol. 51, Part A, pp. 122-130.

Ng, Y. and Pera, P., 2018. Recommending Social-Interactive Games for Adults with Autism Spectrum Disorders (ASD). *Proceedings of the 12th ACM Recommender Systems Conference*. Vancouver, Canada, pp. 209-213.

Quinlan, J. R., (1986). Induction of decision trees. Machine Learning, (1)

Scenicview Academy, 2016. *Bridges We Build: The art of making friends.*

Simpson, R., 2005. Evidence-based Practices and Students with Autism Spectrum Disorders. In *Focus on Autism and Other Developmental Disabilities*, Vol. 20, No. 3, pp. 140-149.

Stevenson, J. and Hart, K., 2017. Psychometric Properties of the Autism-Spectrum Quotient for Assessing Low and High Levels of Autistic Traits in College Students. In *Journal of Autism and Developmental Disorders*, Vol. 47, No. 6, pp. 1838-1853.

VideoGameGeek, https://videogamegeek.com/

White, S. et al, 2009. Anxiety in Children and Adolescents with Autism Spectrum Disorders. In *Clinical Psychology Review*, Vol. 29, No. 3, pp. 216-229.

Wijnhoven, L. et al, 2015. The Effect of the Video Game Mindlight on Anxiety Symptoms in Children with an Autism Spectrum Disorder. In *BMC Psychiatry*, Vol. 15, No. 1, pp. 138.

Wikipedia, https://www.wikipedia.org/

Yerys, B., Bertollo, J., Kenworthy, L., Dawson, G., Marco, E., Schultz, R., & Sikich, L. (2019). Brief Report: Pilot Study of a Novel Interactive Digital Treatment to Improve Cognitive Control in Children with Autism Spectrum Disorder and Co-occurring ADHD symptoms. *Journal of Autism and Developmental Disorders*, *49*(4), 1727-1737.