Comparing Elo, Glicko, IRT, and Bayesian IRT Statistical Models for Educational and Gaming Data

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Educational Statistics and Research Methods

by

Breanna Morrison
Emporia State University
Bachelor of Science in Psychology, 2011
Emporia State University
Master of Science in Psychology, 2013

May 2019
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

_____	_____
Ronna C. Turner, Ph.D. Dissertation Co-chair	Xinya Liang, Ph.D. Dissertation Co-chair
_____	_____
Wen-juo Lo, Ph.D. Committee Member	Samantha Robinson, Ph.D. Committee Member

ABSTRACT

Statistical models used for estimating skill or ability levels often vary by field, however their underlying mathematical models can be very similar. Differences in the underlying models can be due to the need to accommodate data with different underlying formats and structure. As the models from varying fields increase in complexity, their ability to be applied to different types of data may have the ability to increase. Models that are applied to educational or psychological data have advanced to accommodate a wide range of data formats, including increased estimation accuracy with sparsely populated data matrices. Conversely, the field of online gaming has expanded over the last two decades to include the use of more complex statistical models to provide real-time game matching based on ability estimates. It can be useful to see how statistical models from educational and gaming fields compare as different datasets may benefit from different ability estimation procedures. This study compared statistical models typically used in game match making systems (Elo, Glicko) to models used in psychometric modeling (item response theory and Bayesian item response theory) using both simulated data and real data under a variety of conditions. Results indicated that conditions with small numbers of items or matches had the most accurate skill estimates using the Bayesian IRT (item response theory) one-parameter logistic (1PL) model, regardless of whether educational or gaming data were used. This held true for all sample sizes with small numbers of items. However, the Elo and the non-Bayesian IRT 1PL models were close to the Bayesian IRT 1PL model's estimations for both gaming and educational data. While the 2PL models were not shown to be accurate for the gaming study conditions, the IRT 2PL and Bayesian IRT 2PL models outperformed the 1PL models when 2PL educational data were generated with the larger sample size and item

condition. Overall, the Bayesian IRT 1PL model seemed to be the best choice across the smaller sample and match size conditions.

ACKNOWLEDGEMENTS

I would like to thank my Mom for her support as well as my Dad, Sister, and Grandma. I would also like to thank my committee members for having to read this, especially Dr. Turner who had to read it multiple times. Thank you coffee and anyone who has every made me coffee or brought me coffee. Thank you to the people on Stack Overflow who suggested tips to help me when I was stuck on my R code. Thank you also Dr. Robinson for helping me on parts of my code and thank you Dr. Liang for helping me better understand Bayesian statistics. Thank you everyone I have cited in this dissertation who contributed knowledge on the topic and provided a foundation for this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Estimating people's skill or ability levels and using that information to predict outcomes of events are some of the most powerful and practical tools in a statistician's arsenal. Educational researchers use prior performance to predict student academic outcomes (Campbell & Dickson, 1996). Demographic variables can be analyzed to see how they relate to health outcomes in order to determine populations most at risk for certain diseases (Pickett, 2001). Industrial/organizational psychologists can measure personality traits to see how they predict job performance (Tett, Jackson, & Rothstein, 1991). Even video games estimate players' abilities after matches are completed and use that information in predictive models to match players of equal skills for the next game (Véron et al., 2014).

While ability estimation procedures and predictive models are used in many different fields, these fields tend to use different statistical models due to the need to analyze different types of data based on their format, structure, and distributional properties. However, there are times when the underlying mathematical structures of statistical models in different fields are quite similar. One example of this is a comparison of some of the models used for skill estimation in gaming data as compared to ability estimation with educational or psychological data. Elo, and its many variations, are popular models used to match players from a variety of sports and games based on their skills (Coulom, 2007; Glickman, 1995; Hvattum & Arntzen, 2010). The Elo statistical model for estimating a player's expected score from a match is calculated using a logistic function comparing the difference between the player and their opponent's ability ratings. This model is similar to a component of item response theory (IRT), a popular statistical model for the educational field, that estimates a person's expected score on an

item by using a logistic function that compares the difference between a person's estimated ability and an item's difficulty level (Crocker & Algina, 1986; Lynch, 2007; Stocking & Lord, 1983). IRT estimation can be done with frequentists' methods such as maximum likelihood estimation (MLE) or with Bayesian estimation methods which here will be called Bayesian IRT. The mathematical similarities between many of the gaming and educational models are substantial, however the process for applying the models can vary given differences in the structure of the data. A comparison of educational and gaming statistical models with different types of data is the focus of this research. The following is an expansion on the development of the different statistical models that are of interest to this study.

Background of the Study

Elo originated in the 1970s and its simplicity was beneficial when there was a lack of computer processing power and smaller sample sizes were used to calculate player ratings from chess tournaments. There was a need to rank and match players in a way that was standardized and applicable across samples of players. Elo used matched comparisons to take into account not only the outcome of the game but the skill differences between the players and it could be calculated quickly, even by hand (Elo, 1978). Elo and its variations are still popular even when the disadvantage of low computer processing power is no longer a concern such as when using online game matching ("How does GameKnot's rating system work?," 2017). Elo is a relatively simple set of mathematical formulas where performance ratings are based on the number of wins and losses which are then weighted by the player's opponents' ratings. The Glicko formula includes an additional parameter that is presented as a standard deviation which takes into account the estimated accuracy of the players' ratings (Glickman, 1995). Many gaming systems still rely on the use of Elo and Glicko procedures to estimate player ability, and to pair players

for online and in-person tournament games. The use of Elo can be beneficial for ranking tournament members in circumstances where the number of players is relatively small as there are not as many calculations to compute. However, online gaming sites are no longer limited to small sample sizes or lack of processing power, therefore it may be useful to explore how more advanced statistical models such as IRT and Bayesian IRT may function with larger, more complex gaming data. Elo and Glicko are the two gaming ability estimation models that are the focus of this study.

For ability estimation models that pertain to educational related data, IRT has been a popular model used to evaluate test items and how responses to those items can relate to a person's ability level. It was a way to measure difficulties of individual items allowing those items to be used in different test batteries and with populations of varying ability levels (Baker, 2001). Another statistical model for IRT estimation whose popularity grew with the increasing efficiency of computers is Bayesian IRT. While the idea of Bayesian statistics has been around since Bayes' theorem in the 1700s, Bayesian statistics has only recently grown in popularity, becoming relevant in a wide variety of fields (Lynch, 2007). The Bayesian approach allows for the use of prior information to be included in estimations based on a sample. That is, instead of relying only on the sample data, prior values can be used as a starting point in estimation. Bayesian approaches also seek to model the whole distribution of the data rather than just the best point estimate. The use of Bayesian estimation for IRT models helps improve the convergence and allows for a smaller sample size needed than for IRT if the prior assigned is relatively accurate (Gelman et al., 2014).

In practice, when teachers wanted to look at ability estimations, they tend to use proportion correct due to smaller sample sizes available to them and possibly limited statistical

experience. Proportion correct is how many questions out of the total a participant answered correctly (Hambleton, Swaminathan, & Rogers, 1991). However, proportion correct is limited to only being appropriate for comparisons when a sample has completed the same set of items.

All of these models seek to estimate people's skill levels or abilities and predict the chances of success on an outcome whether that outcome is getting an answer right on a test or winning a chess match. With increases in computer technology, a wider variety of complex statistical models are available to populations of gaming groups and educational professionals. Some of the more complex statistical models could assist gaming statisticians in obtaining more accurate ability estimates for player matching within large datasets. Conversely, researchers with smaller, less comprehensive datasets may benefit from the use of statistical models that have fewer model assumptions for evaluating student performance. However, statistical models do not always function effectively or accurately under all conditions. Thus, knowledge of how statistical models from fields such as education function for gaming data and how gaming statistical models function for educational data can assist researchers and practitioners in selecting which models to use for their data.

While each model has its advantages and disadvantages, there have not been many studies investigating how these fields can benefit from each other's models. Some of the studies that have compared Elo and IRT have had methodological limitations (Antal, 2016; Pelánek, 2014; Wauters et al., 2012). For instance, when using a real dataset, researchers used IRT estimations on the whole data set as the "true" ability for comparing to Elo estimations (Antal, 2016; Wauters et al., 2012). This makes the assumption that IRT was the best estimate under all conditions which may not have been true.

Further, the type of data analyzed can have an impact on the accuracy of the models. When simulations are conducted (Pelánek, 2014), a model must be selected for generating the data. If data are generated using an IRT 1PL model, it might be reasonable to expect that an IRT 1PL model might be more effective in estimating the ability levels of persons within that dataset as compared to other types of models such as Elo or Glicko. Similarly, if data are generated using Elo models, an Elo statistical model might be expected to best estimate the sets of response strings. The above studies that compared Elo and IRT models (e.g., Antal, 2016; Wauters et al., 2012) used educational data to compare these models and little was found that investigated how IRT models perform on gaming data. It would be of interest to investigate whether IRT and Bayesian IRT models can provide accurate estimations with gaming types of data and to further what little research has been done on using gaming statistical models on educational data.

Study Overview and Research Questions

According to the literature review, this dissertation aims at comparing how accurately two gaming models (Elo and Glicko) and five models commonly used in education (proportion correct, IRT 1PL/2PL, and Bayesian IRT 1P/2PL) were able to estimate ability values for gaming and educational data. The different assumptions required for gaming and educational statistical models and the changing fields of education and gaming research may lead to statistical models typically used in one field being beneficial in the other field under some circumstances.

The specific research questions are as follows:

1. How do estimates of ability using gaming and educational achievement statistical models such as proportion correct, Elo, Glicko, IRT 1PL/2PL, and Bayesian IRT

1PL/2PL, correlate with true ability using simulated gaming and educational achievement type data under varying conditions?

   a. Which ability estimates are most correlated to true ability simulated using a gaming data model?
   b. Which ability estimates are most correlated to true ability simulated using an educational achievement data model?
   c. Which ability estimates are most correlated to true ability when the number of items and game matches vary from small, moderate, to large?
   d. Which ability estimates are most correlated to true ability when sample size varies from small to moderate?
   e. Does data generation method influence the correlations between model estimations and true ability?
   f. Which ability estimates produce the smallest standard errors in relation to true ability estimates for the gaming and educational achievement data under the varying item/match size and sample size conditions?

2. Which statistical models (e.g., gaming models such as Elo, Glicko; educational models such as proportion correct, IRT 1PL/2PL, Bayesian IRT 1PL/2PL) best predict real data outcomes?

   a. Which ability estimates best predict real outcomes for certain number of matches for gaming data?
   b. Which ability estimates best predict real outcomes for certain items in educational data?

The sections of research question one were investigated through a simulation study, and the sections for research question two were investigated using empirical studies. In the simulation study, four types of simulated data (Elo, Glicko, IRT 1PL, and IRT 2PL) were used to control for differences in model estimations that may occur due to data generation methods. Other manipulated conditions include varying the number of participants (50 and 150) and the number of matches/items (5, 15, and 30). Various sample sizes were considered because iterative estimations used in IRT and Bayesian IRT may not be feasible for small sample sizes (Foley, 2010; Şahin & Anıl, 2017), while Elo and Glicko may be able to estimate abilities under smaller sample size conditions. Additionally, if a gaming dataset has a large enough sample size, it may benefit from more advanced models such as IRT as compared to the more simplistic Elo and Glicko models.

While simulations allow for the true ability to be known, generating the data requires a decision on a formula in which to generate the data which could bias results. Thus, an empirical data section was also included. There were two empirical datasets, one being a dataset of online chess matches and the other educational achievement test items from the Trends in International Mathematics and Science Study (TIMSS). This allowed for two different types of datasets to be evaluated with the statistical models. While the correlations between the ability estimations were calculated, a small sample of data was set aside and used to evaluate the predictive power of the estimations as a way to obtain a measure that investigated whether certain models have greater predictive accuracy under certain conditions. Having both a simulation part of the study and an empirical data part of the study should better show trends regarding the statistical models.

Not only would this study provide support for the usefulness of comparing statistical models from the educational and gaming fields, it may help show the importance of

understanding statistical models from different fields and encourage comparisons of statistical models from fields beyond just gaming and educational models. This study will help provide guidelines for the advantages and disadvantages of these models under the situations studied.

CHAPTER 2

LITERATURE REVIEW

This literature review will describe a selection of popular ability estimation models for both gaming data and social science data. There will be a focus on Elo, Glicko, IRT 1PL, IRT 2PL, Bayesian IRT 1PL, and Bayesian IRT 2PL models and some related models that fit into those categories. Advantages and disadvantages of each will be discussed, and literature that has explored how these models compare to each other will be reviewed.

Statistical Models for Gaming Data

Statistical models for estimating ability in gaming are often used for matching players together for close matches or for qualifications for tournament purposes. Players are also interested in quantifiable measures of their own skill. New estimations can be generated after every game or after a set of games (Glickman, 1995). While there are many different gaming estimations models, the focus of this dissertation will be on Elo and Glicko.

Elo

Uses and origins. The Elo rating system was originally used for chess and named after its creator Arpad Elo (Elo, 1978). It was used as a system to assign chess players' skill levels to adequately match them with players of similar skill. While its origins lie in chess, the Elo rating system, and its related Bradley-Terry model, has been expanded to other games such as other board games, football, tennis, and video games (Coulom, 2007; Glickman, 1999; Lasek, Szlávik, & Bhulai, 2013; Véron et al., 2014).

Rating the skill of chess players happened before the implementation of the Elo system, however there were some issues with these rating systems. Previous chess rating procedures were more subjective. In fact, it was possible for chess players to lose games and still gain ability

rating points. Other systems used in game skill ratings, such as the American Contract Bridge League rating system, focused more on a ladder rating where skill rating can only go up but cannot adjust back down (Glickman, 1995). The Elo model is typically a purely statistical based formula where skill rating is computed based on the aggregated wins and losses of players weighted by how likely the win or loss was based on the opponent's ability. Some have even proposed that the adoption of the more objective and useful Elo ratings by the International Chess Federation in 1970 may have been one of the key factors of increasing the popularity of chess tournaments (Glickman, 1995).

The uses of Elo in chess have varied purposes. It can be used to match players against each other as well as using skill rating to have cut-offs for different skill tiered tournaments. Some systems have titles earned in chess tied to Elo ratings. Elo uses in chess may differ from other game systems. For instance, tournaments often avoid pairing high skill ranking players together early in the tournament and may structure the tournament match in a way to most likely allow the "big players" to make it to the finals in order to make the finals more appealing and interesting to watch (Glickman, 1995).

Elo uses in other game systems may be more automated. For instance, Elo related ranking systems are popular in multiplayer games. Players in multiplayer games are often placed in a queue while the system uses a variety of measures such as ping, time already spent waiting, and Elo rating to match players together (Véron et al., 2014).

While Elo is a popular way to rank players in a variety of games there has been some application of Elo to other fields as well. Some researchers developed Elo based models in predicting animal behavior while others used Elo for detecting deficiencies in fabric patterns (Newton-Fisher, 2017; Tsang, Ngan, & Pang, 2016). Some researchers have also advocated

some benefits of using Elo type formulas when evaluating student performance (Brinkhuis & Maris, 2009; Pelánek, 2016).

Formulation. The Elo formula is a relatively simple but malleable way to rank skill. Its main premise is that a win will increase one's skill rating while a loss will decrease one's skill rating. Additionally, how much a person's skill rating changes is a function of their calculated probability of winning the match which is based on their opponent's skill rating. If their probability of winning is around 50%, the change in the player's skill rating is relatively equal in either the positive (winning) or negative (losing) direction. However if a player with a much lower skill rating is competing against a much higher rated player, the lower skilled player will gain a higher amount to their skill rating for a win than the higher skilled player to account for the lower skilled player's lower probability of winning (Elo, 1978).

The way an initial Elo rating is calculated can differ. The World Chess Federation offers methods to calculate an unrated player by averaging the skill ratings of fellow players competing in the tournament ("FIDE Rating Regulations," 2014). Online chess game sites may also use a different estimation system for the first twenty games before moving to an Elo way of adjusting skill rating ("How does GameKnot's rating system work?," 2017). Online gaming often assigns an unrated player an average skill rating and then adjusts from there using data obtained from the player's activity (Véron et al., 2014).

Assuming there is an initial skill rating given to the player (even if it is an "average" rating), a probability of the player's chances of winning can be calculated when given the skill rating of their opponent. The probability is a logistic function with the typical feature of the probabilities of a win less than 100% even when a very high skill player is paired with a very low skilled one. Subsequently the chances of a win can never be 0%. If *P* represents the player's

current skill rating and *O* is the opponent's skill rating, then one calculates *E*, the expected probability of winning, by:

$$E = \frac{1}{1+10^{-(P-O)/400}} \tag{2.1}$$

The numbers 10 and 400 are additional parameters that represent the scaling of the skill rating being used. The scalers used here relate to the skill rating used for chess which ranges between 1000 and 3000 with an average of around 1500, although the Elo scale can be changed (Elo, 1978). The probability of an expected win is then used when calculating the change in skill rating to be applied to a person's current skill estimate. To calculate the new skill rating (*New*), let *P* represent the current skill rating, let *Out* be the outcome of the match (1 for a win and 0 for a loss), and let *E* be the probability of winning the match and K is a constant, frequently 32 in chess. The formula for the new skill rating is:

$$New = P + K(Out - E) \tag{2.2}$$

The above formula shows that the more unexpected the outcome of a match (the greater the difference between *Out* and *E*), the larger the adjustment to the player's rating. When a skill rating can change also differs. In official tournaments a skill rating is often only updated after all of the games of the tournament are completed. This alters the formula making Out change to the average of outcomes and *E* the average probabilities of winning ("FIDE Rating Regulations," 2014). Online gaming systems tend to update skill ranking after each game ("How does GameKnot's rating system work?," 2017; Véron et al., 2014).

Mathematically, a group of players whose skills are rated using Elo will end up creating a normal distribution of skill ratings. While purely using the Elo formula would mathematically result in a normal distribution, the World Chess Federation sets the minimum Elo to be 1,000, artificially creating a floor effect in the distribution. Players whose skill ratings would

mathematically be lower than 1,000 are considered unranked ("FIDE Rating Regulations," 2014). While the US chess rating system also has a minimum score set, for the US system the lowest score is 100 instead of 1,000 allowing for a more normal distribution (Glickman & Doan, 2017). The USCF (United States Chess Federation) system has also implemented other restrictions, such as having a skill rating not fall 200 points below their highest ranking. This was implemented for a multitude of reasons such as to keep players from manipulating the system in order to participate in lower skill tiered tournaments and to keep players encouraged instead of having their skill rating drop too much (Glickman & Jones, 1999).

While the above shows the most basic Elo formula, there have been a variety of formulas that alter it. A popular alteration is to change *K* from a constant to a number that is based on the number of games completed. The US chess rating system employs a *K* that is not a constant but instead is calculated as:

$$K = \frac{800}{N + m}, \tag{2.3}$$

where N is the total number of effective games that have been completed and m is the number of games in the current tournament that was played. "Effective" games is a measure partly depending on the number of games the player has completed or is set at a max of 50 (Glickman, 1995). Another possible alteration is to change the static scaling variable in the probability formula into a moving one. This procedure was adopted due to findings that the expected probabilities were biased, that is, they overestimated the probability of a win for very high and very low skilled players (Glickman & Jones, 1999). The formula for this varying E value is:

$$E = \frac{1}{1 + 10^{-(P-O)/a}}, \tag{2.4}$$

where *a* is a non-constant number (Glickman & Jones, 1999). For instance, one study found that changing *a* from 400 to 561 in chess estimates, improved the fit of the model for the data being used. They used maximum likelihood methods to find the parameters with the best predictive power and then worked backwards from there to determine what value *a* should be (Glickman & Jones, 1999). The value of *a* changed depending on the skill ranking of the player and helped with the issue that the winning probability for high rank players was occasionally overestimated by the formula.

Glicko

Another variation on Elo is the Glicko formula. The Glicko formula seeks to improve estimation accuracy by having a previously fixed constant in the formula be an adjusted value (Glickman & Jones, 1999). The Glicko formula adds a standard deviation to the player's ratings in order to give more information about the certainty of the rating. This is called the Rating Deviation (RD) and is shown here as RD_{PNew}. To calculate the RD for a player:

$$RD_{PNew} = \sqrt{(RD_p^2 + c^2 * t)}\,, \tag{2.5}$$

where RD_p is the current player rating (a new player would get a constant decided upon), c is a constant based on how malleable RD will be over time, and t is based on how long it has been since the player has played. The above formula is used with a new rating period, that is after a certain amount of time has passed between ratings such as updating RD from one tournament to another. The formula for calculating RD between matches is different and will be shown below. For match by match Glicko calculations, the probability of a win is very similar to the Elo formula but with a few additions:

$$E = \frac{1}{1 + 10^{-g(P-O)/a}} \tag{2.6}$$

The parameter *a* can be set to 400 like in the previous Elo formula or it can be changed. Here it will mirror the Elo value for our study. The new parameter *g* is added which is calculated using the opponent's RD (RD_O):

$$g = \frac{1}{\sqrt{(1 + 3q^2 * \frac{{RD_O}^2}{\pi^2})}}, \tag{2.7}$$

and q is a constant such as:

$$q = \frac{ln10}{400} = .0057565 \tag{2.8}$$

The RD for the player is also updated:

$$RD_{PNew} = \frac{1}{\sqrt{\frac{1}{{RD_P}^2} + 1/d^2}}, \tag{2.9}$$

where d^2 is:

$$d^2 = \frac{1}{q^2 * g^2 * E(1 - E)} \tag{2.10}$$

That just leaves the player rating update formula which again is very similar to Elo but with added parameters:

$$New = P + qRD_{PNew}^2 g(Out - E) \tag{2.11}$$

Basically, the Glicko adds a standard deviation to the ratings to allow for the certainty of a player's rating to factor into adjustments made. A player's rating will change more if their opponent's RD is smaller indicating more certainty that the opponent is accurately ranked (Glickman, 1995). Additionally, if enough time has passed between matches, formula 2.5 will be used to further update RD. There have even been updates to this formula that add a parameter that measures true variability in the player's ratings. That is, it is a standard deviation that is not

based on measurement error but on how erratic the player's performances are which is called volatility (Glickman, 2001).

Gaming Statistical Models Application

Like any statistical model, Elo and Glicko have advantages and disadvantages. One of the main benefits of Elo and Glicko are their simplicity. Elo was used before computers were readily available, so having a simple formula allowed people to easily calculate and update skill ranking. Even with computers, there is a benefit for more simple formulas. Multiplayer games update skill rankings after every game. A formula that relies on running iterations in order to calculate a new skill ranking may take too long to justify its use when gamers do not want to wait more than a minute to be matched in a game (Véron et al., 2014; Weng & Lin, 2011). The Elo and Glicko formulas do not have that disadvantage making it preferential when one seeks to find a quick calculation that avoids using heavy computer resources.

There are also assumptions allowed in Elo and Glicko that might not apply to other statistical models. Elo and Glicko allow for the assumption that a person's skill rating can and will change (Elo, 1978). Many measurement models discussed later have the assumption that the skill being measured will not change over the time of the testing. Further, some measurement models, such as some IRT testing, occasionally use additional test and item information to measure one's skills (e.g., item parameter estimations from prior samples). This is especially prevalent when IRT is used for computer adaptive testing which is when participants are matched to items based on their ability levels and the item's difficulty (Baker, 2001). When using Elo and Glicko to match online players, a system similar to computer adaptive testing, there is little need for any testing to be completed beforehand as skills are continuously being estimated, updated, and matched. Even when someone is being compared to an unranked player,

the unranked player can be assigned a starting number which may or may not be based on previous information such as the average rating of players (Glickman, 1995).

There have been attempts to adjust the basic Elo formula to address some of its disadvantages. The basic Elo formula has demonstrated that its prediction accuracy varies based on skill level and number of games. One possible limitation could be the lack of a variance estimator in the Elo statistical model for individuals which the Glicko model attempts to provide.

Another disadvantage to Elo and Glicko formulas is that researchers tend to subjectively alter the use of the formula in order to get a more preferred distribution. Skill rating floors are used in both the US Chess Federation and the International Chess Federation ("FIDE Rating Regulations," 2014; Glickman & Doan, 2017), and the value of the *a* parameter is adjusted for certain conditions. Uses of the Elo formula that apply subjective bonus points in order to keep the average of the rankings stable may also be a disadvantage since it allows for subjectivity in the assigned ability values (Glickman & Jones, 1999). Even the pure Elo formula is subjected to bias while more standard social science measurement models tend to be unbiased (Brinkhuis & Maris, 2009).

There are also data that are lost that could be potentially used. The Elo and Glicko systems typically rate a win as 1, a draw as .5, and a loss as 0 (Elo, 1978). A close win and a clear win provide different information about the player's skill rankings but each are considered equal in the Elo and Glicko formulas. Further, information can also be obtained from a draw or a loss with not all draws and all losses being equal though the formula treats them as such.

Like any paired comparison calculation, there are issues when the comparisons are made within only a certain pool of people. This is frequent in chess. A player who only competes in junior tournaments is getting their skill rating calculated with only other junior players. This

leads to their skill rating being inaccurate when they move to adult tournaments (Glickman & Jones, 1999). While this can be fixed by having everyone have an equal chance of competing against everyone else rather than grouping players, that solution may not be feasible. Even for online games, systems often rely on pairing players by location in order to prevent connection issues (Véron et al., 2014), and pairing players of closer ability estimates to make the games more interesting (and less discouraging) for the participants.

Another limitation of Elo and Glicko ratings is that official ratings tend to not be updated often unless the chess player is active in many tournaments. If a chess player has played in one tournament then spent a few years just playing with friends, their skill ranking would officially still be what their tournament rating was while their true skill rating may be much higher due to having additional practice that has not been accounted for in official measures (Glickman & Jones, 1999).

Elo and Glicko also assumes that it is one player versus another player. In team-based games this assumption is not met. While there are people who treat the whole team as an individual or sum the skills of players within a team, the skill ranking results from such uses tend to not fit as well as they should (Herbrich & Minka, 2007; Lasek et al., 2013). As this study focused on estimating abilities for individuals, this disadvantage was not a factor however there will be other gaming statistical models that will be mentioned later that do not have this disadvantage.

Changes in Elo skill ratings when the players have vastly different skill rankings poses its own set of challenges in both a practical and statistical sense. For instance, a player with an Elo of 2382 has a 90% chance of winning against a player with an Elo of 2000. However, if they played ten games and, as expected, the higher-level Elo player loses one game of those ten, that

player actually ends up losing two skill points due to rounding and updating the skill ranking at the end of all matches as tournaments often do. This estimation difference gets worse if the ratings are updated after each match. If the higher skill player wins the first 9 games then loses the 10th, they have a net loss of 5 points while the lower skilled player has a net gain of 5 points. There are also practical implications when players of vastly differing skills are matched as one player is much more invested in avoiding a loss while the other is more invested in a win.

With both benefits and disadvantages to Elo and Glicko statistical models, it is useful to see how these models compare to other statistical models. When looking at how Elo works, I was reminded of IRT models. In Elo for gaming purposes, a player is matched over and over again based on their skill ranking while adjustments are made based on the outcomes of the matches. This reminded me of the way a person gets matched with item after item in order to obtain and estimate their skill rating as is the case for some computer adaptive testing (CAT) systems which may use IRT to select from item pools of varying difficulties. There have been some comparisons of the two models already and more detail regarding statistical models for educational data will be discussed (Antal, 2016; Pelánek, 2014; Wauters, Desmet, & Noortgate, 2011; Wauters, Desmet, & Van Den Noortgate, 2012).

Statistical Models for Educational Data

There are many ways to measure ability using educational data. One popular model is IRT. There are also Bayesian estimation methods that can be used with IRT which is called Bayesian IRT. One of the simplest ways of estimating ability is calculating the percentage of test items a participant has gotten correct known as proportion correct (Baker, 2001). These models, used in the field of educational and the social sciences, were selected for this study.

IRT

Uses and origins. Item Response Theory (IRT), has offered a way to improve upon traditional educational and psychological testing measures offered by classical test theory. Proportion correct is a way to estimate ability levels using classical test theory and it is calculated by taking the number of items correct divided by the total number of items administered. While its simplicity gives it some advantages, proportion correct is very limited. People's abilities that are estimated with proportion correct are only comparable to people who have taken the same test or an equivalent one. Additionally, if one wanted to use an iterative matching system, like many computer tests and game matching systems do, proportion correct would not function well as its ability estimations are highly test dependent. Due to these disadvantages, IRT was explored as a way to evaluate item level difficulties independent of test composition (Crocker & Algina, 1986).

Similar to Elo, there was a history of the statistics behind the IRT model before the model started gaining widespread use. Some of the works of Thurstone and Binet in the early 1900s relating to cognitive testing resembles some of the components of IRT models (Bock, 1997). However, the big growth in interest in IRT may be linked to *Statistical Theories of Mental Test Scores* (Lord & Novick, 1968), a text published in 1968 that combined works from previous researchers' publications relating to IRT and its various models. That along with an increase in computational power helped lead the way to more computationally extensive statistical solutions like IRT (Jones & Thissen, 2006).

Both its origins and current uses have been heavily focused on measuring cognitive ability (van der Linden & Hambleton, 1997). However, there are also IRT models focused on other types of measurements such as personality traits (e.g., Gray-Little, Williams, & Hancock,

1997; Steinberg & Thissen, 1996). Additionally, like Elo, there are multiple formulas and variations of IRT.

One main goal of IRT is to categorize questions used in a measurement tool into differing levels of difficulty. Each item has an underlying distribution that shows the likelihood of a correct answer depending on a person's skill level. These items are then used to analyze a participant's knowledge or skill through a process of administering items to participants and evaluating their performance. Through iterations, the most likely skill level is calculated for participants based on their correct and incorrect responses to the set of items (Hambleton et al., 1991).

IRT theory is very popular in tests that use computer adaptive testing (CAT). Computer adaptive testing is a process that will estimate a person's skill level by giving them items with a difficulty that will correspond to their skill level or ability. For instance, if a participant misses three average difficulty level questions, the computer will start giving questions of lower difficulty. If a person answered all three correctly, the computer will give them more difficult problems (Hambleton et al., 1991). In this way, you can view the participant and the item as a paired comparison, similar to Elo or Glicko.

Formulation. While there are many IRT related formulas, the three most notable models for dichotomous data are the 1PL model (the one-parameter logistic model, also known as the Rasch model), the 2PL model, and the 3PL model. For the 1PL model (Baker, 2001), the only parameter allowed to vary for calculating the probability of getting a question right is item difficulty (*b*). Item difficulty is defined as the skill level needed in order to have a 50% probability of getting the correct answer on the item. Item discrimination is either set to a value of 1 when using the Rasch model, or all items are given the same discrimination value (not

necessarily equal to 1) for the 1PL model (van der Linden & Hambleton, 1997). If *b* is item difficulty and θ is the person's ability level, then the formula to calculate the probability of a person with a certain ability getting a question right, *P*(θ), using the 1PL model with item discrimination set to 1 would be:

$$P(\theta) = \frac{1}{1 + e^{-1(\theta - b)}} \tag{2.12}$$

The term *e*, Euler's number, is a constant equivalent to approximately 2.71828. If a person has average skill (θ = 0) and they answer a question with a difficulty of 0 then they would have a 50% probability of answering the question correctly. The ability ratings are calculated with a mean of 0 and a standard deviation of 1.

The 2PL model adds an additional parameter *a* that represents item discrimination (Hambleton et al., 1991). Item discrimination is how well an item discriminates between people of differing skill levels. If the probability of someone with average skill answering a question correct is similar to someone with a much higher skill level answering that same question correct, then the item would have low item discrimination. If two participants with a small ability difference have very different probabilities of answering a question correct, then it would be an item with high item discrimination. The 2PL model allows *a* to vary where the probability of answering the question correctly given a certain skill level would now be:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} \tag{2.13}$$

The 3PL model adds one more parameter that is based on guessing. While an item that is "fill in the blank" may have a near 0% chance of getting the item right by guessing, there is a substantial chance of guessing the right answer when the question is multiple choice. If *c* is the

guessing parameter, which represents the chance of getting the question right if the person had a skill level nearing negative infinity, then the 3PL model would be:

$$P(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}} \tag{2.14}$$

The item characteristic curve (ICC) is an essential aspect of IRT representing the probabilities of answering an item correctly among a variety of skill levels and can help demonstrate the three parameters discussed. The ICC is shown in Figure 1. The *b* parameter of item difficulty shows that a theta of -.21 is the point where a participant has a 50% chance of answering an item correctly. The *a* discrimination parameter is the slope of the curve. Here, the *c* parameter is zero, and this is shown by having the lower left tail of the ICC approaching zero. The figure shows people with lower skill levels have a lower probability of answering the question correctly, while someone with a higher skill level has a higher chance of answering the item correctly (Baker, 2001).

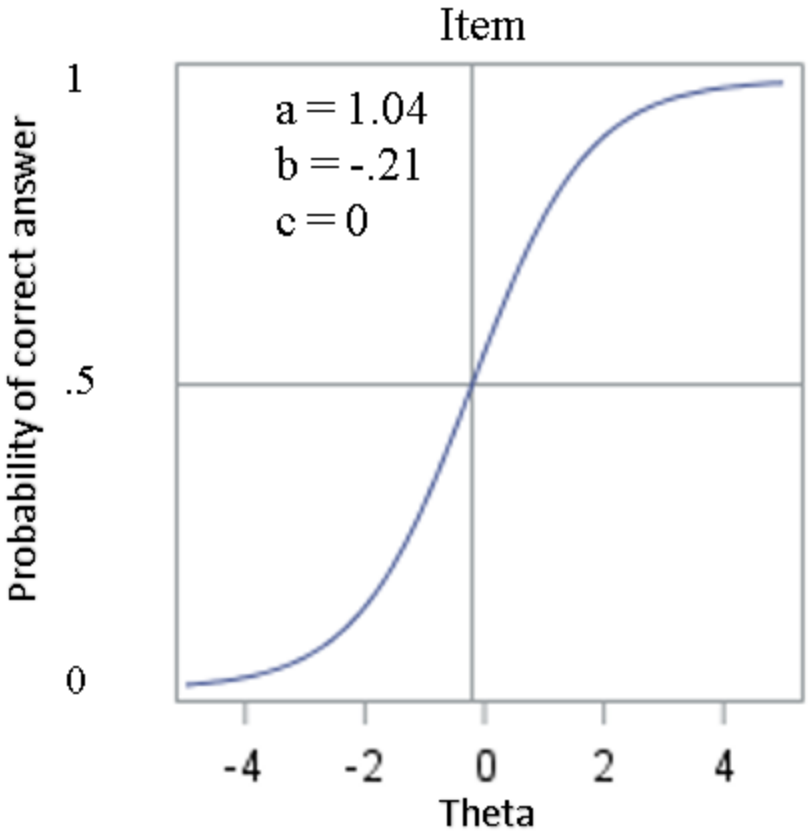

Figure 1. 3PL IRT model demonstration

IRT calculates participants' ability estimations based on their responses to a set of items. IRT uses individuals' response strings to estimate both item parameters and the most likely theta or ability estimate for each person based on their response string. The formula for estimating ability using a 2PL model would be (Baker, 2001):

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \frac{\sum_{i=1}^{N} -a_i[u_i - P_i(\hat{\theta}_t)]}{\sum_{i=1}^{N} a_i^2 P_i(\hat{\theta}_t) Q_i(\hat{\theta}_t)} \quad (2.15)$$

Estimated ability for the participant is represented by $\hat{\theta}$ with the *t* representing the ability estimated in the prior iteration. The first iteration usually puts the person at either an average ability or a value based on number correct. Then, based on that ability and their response string,

their new ability estimation is adjusted as represented by $\hat{\theta}_{t+1}$. The recommended change in ability is calculated by summing across differences in item score (u_i = 1 for correct versus u_i = 0 for incorrect) and the probability of getting a correct response for an item *i* given the participant's estimated theta ($P_i(\hat{\theta}_t)$, weighted by the discrimination of item *i* (a_i), divided by the sum of the products of the probability of getting an item correct given theta, the probability of getting an item incorrect given theta ($Q_i(\hat{\theta}_t)$), and the squared discrimination of the item. Ability estimations are completed iteratively along with item parameter estimates until changes in estimations meet a minimum threshold. Overall, the closer the calculated probabilities for the current theta are to the actual outcomes in the set of items, the smaller the adjustment is made in theta (Baker, 2001).

For estimating the item parameters in IRT, a maximum weighted likelihood function is used. This is a typical maximum likelihood value where parameters are set to maximize the likelihood function, a function that shows the probability of the parameters with the current data. The likelihood function for the 2PL model is:

$$L(\boldsymbol{u}|\theta, b, a)\,, \tag{2.16}$$

where *L* indicates the likelihood function and $\boldsymbol{u}$ is a vector of responses to items from a participant. The determination of these parameters is calculated by multiple derivatives with starting values of the parameters based on the parameters that could be obtained from classical test theory. Then estimations are completed in an iterative process that results in increasingly better fitting parameters if the data meet the statistical model's assumptions. The parameter estimation process is shown by:

$$\mathbf{x}_i^{j+1} = x_i^j - \left\{H\left[x_i^j\right]\right\}^{-1} f'\left[x_i^j\right], \tag{2.17}$$

where vector **x** is representing the estimation of the two parameters in a 2PL model (Hambleton, 1991). The iteration number is represented as *j* and the *j+1* means it is over iterations with *j* being the current iteration. The matrix of the second derivatives (noted by $\{H[x_i^j]\}$) is derived from the matrix of the first derivatives ($f'[x_i^j]$). Item parameter values are determined at the point the change in the estimation reaches a sufficiently small value (Hambleton et al., 1991).

As mentioned before, there are many IRT models. The 1PL, 2PL, and 3PL models that assume dichotomous outcomes have been presented. The graded response model assumes that the answer is not categorized by just a correct or incorrect answer but could have a polytomous outcome that could include partial credit (having varying degrees of correctness). An example would be an essay question where participants can earn a range of points rather than just being marked as correct or incorrect (Hambleton et al., 1991). This model is also used for measurements that use Likert type scales that do not assume a right or wrong answer (Gray-Little et al., 1997; Steinberg & Thissen, 1996). However, since this dissertation is focused on comparing measurement models used in games to more academic statistical models used for ability testing which are commonly binary, the polytomous-based models were not explored. Further, since the gaming data of focus is on paired comparisons between players, the guessing parameter in the 3PL model is not as relevant, therefore the focus of this study was on IRT 1PL and 2PL models.

IRT models application. Like Elo and Glicko, there are both advantages and disadvantages concerning the use of IRT. One of the most important advantages of IRT, as it compares to Elo, may be its increased accuracy. Not only does the IRT model allow for more parameters to be estimated, thus allowing for more accuracy in the model if those parameters are present in the dataset, IRT estimates parameters using all the data available at once while Elo relies on adjusting initial estimates as new data appear. Studies looking at how IRT compares to Elo often use the estimates of IRT as the most accurate model to compare with (Pelánek, 2014; Wauters et al., 2012). This increased accuracy comes at the cost of computational power. The maximum likelihood estimate (MLE) often used in IRT can be slow and strenuous (Wauters et al., 2012). Maximum weighted likelihood estimation can be a more accurate method of MLE and was used for this study, but similar to MLE it takes just as much time (Warm, 1989). For many games that estimate a new skill ranking after every match, using a formula that takes a comparably long time to calculate may not be reasonable especially when it needs to be completed with thousands of players at a time.

An advantage of IRT models over statistical models for gaming data is that, while Elo suffers from issues where skill rankings are largely dependent on the people the player surrounds themselves with, IRT items can work well on other pools of people the items were not initially tested on. Often in IRT, the items used for evaluating a person's ability are tested on a large sample of people that represent a wide range of skills. Subsequently, this makes the participant's ability rating more comparable to others even if they are from different groups (Baker, 2001). However, even when item parameters are estimated from a group with an ability distribution that is different from the group to which the items will be applied, indeterminacy is built into the model such that the new estimates can be rescaled based on differences in estimated difficulty of

the items or matches (Wauters et al., 2012). While statisticians can focus on getting a varied sample in which to use Elo estimates, the researchers using Elo in practice may not be as concerned with the generalizability of their Elo estimates across samples. As mentioned, when Elo is used for tournament rankings, those estimations may not be applicable to the player if they join other tournaments potentially with a different population whereas the IRT estimates may allow for better cross-tournament comparisons.

However, this advantage comes at a cost. There is extensive effort into testing items before they are used to estimate a participant's ability for computer adaptive testing. Elo does not rely on preliminary testing, with skill ranking being calculated when players start completing an adequate number of matches or even after their first match.

IRT also requires a large sample size in order to estimate item parameters with sufficient accuracy, and there are increased sample requirements for more complex models. Elo can estimate parameters with small samples as it uses a more simple paired comparison approach rather than more advanced estimation procedures for parameters (Elo, 1978; Hambleton et al., 1991). There have been studies looking at how IRT works with smaller sample sizes. Foley's (2010) study using IRT with small sample sizes found that while sample size had minimal impact on the correlations between true and estimated abilities, sample size affected the accuracy of the item parameters. Though this particular study was looking at an augmentation technique that could improve IRT item parameter estimations with smaller sample sizes, it did show that even with moderate sample sizes (250) and item sizes (30), the correlations between estimated and true ability were usually near .9. However, the root mean square errors (RMSEs) were much larger than desired, being around .45 for the same condition (Foley, 2010).

Another disadvantage may lie in the assumptions of IRT. IRT assumes that only one unidimensional composite of skills is being measured with the 1PL, 2PL, and 3PL models (Hambleton et al., 1991). For instance, a geometry question should primarily be measuring skills related to geometry. If the question has difficult vocabulary in it, then there might be an issue with the question now measuring multiple skills, one language based and one math based. The assumption that only one unidimensional skillset is related to the outcome of the question is not an assumption Elo makes. The lack of this assumption may be useful when testing for skills that could possibly be multidimensional. It is reasonable to infer that many games might take into account a variety of skills meaning that the ability estimate for a particular game might not be unidimensional.

IRT also has the assumption of local independence. This is related to how the answer to one question should not influence the probability of answering another question correct. This assumption is also related to IRT not assuming that a person's skill or ability can change during the process of measurement. Because IRT is used for measurements that usually occur at a singular point in time, the assumption that skill does not change over the period of time is reasonable (Hambleton et al., 1991). However the process of estimating a player's skill ranking in games is spread out over a much longer period of time and Elo assumes that skill will change (Glickman & Jones, 1999). If one wanted to alter IRT to estimate a player's skill ranking, this violated assumption would be a limitation. The different advantages and disadvantages of both IRT and Elo lead to a growing alternative in skill rating calculations across a variety of fields in the form of increasing the use of Bayesian statistics.

Bayesian IRT

Bayesian Statistics. The history of Bayesian statistics is a long and interesting one. Reverend Thomas Bayes, credited with being the founder of Bayesian statistics, wanted to answer a basic question of how to predict the probability of future events based on past events. The idea is that new data can be used to improve initial prediction (McGrayne, 2012).

Pierre-Simon Laplace later expanded upon the idea of Bayesian statistics with a basic formula focused on more concrete details about calculating updated probabilities based on previous knowledge. However early criticisms of Bayesian statistics viewed the model as too subjective, a criticism that followed Bayesian statistics throughout the next few centuries. Bayesian statistics came back into favor when it was used in WWII for analyzing codes and was also used in the 1950s for medical research (McGrayne, 2012).

Today the uses of the Bayesian approach are numerous. It is a popular estimation procedure for models in the healthcare field and social sciences, such as education, and has even been used in predicting presidential elections (Linzer, 2013; Lynch, 2007; Spiegelhalter, Abrams, & Myles, 2004). Additionally, there has been a growing interest in using the Bayesian approach for predicting gaming outcomes (Coulom, 2008; Herbrich et al., 2007; Weng & Lin, 2011).

Formulation. The most basic Bayesian formula, known as Bayes' Theorem, solves for the probability that a hypothesis (H) is true given that we have certain data (D), otherwise known as a posterior probability (Brewer, 2009). Let likelihood be represented as $P(D|H)$, marginal likelihood be $P(D)$, prior probability be $P(H)$, and the posterior probability be $P(H|D)$ then Bayes' theorem is:

$$P(H|D) = \frac{P(H) * P(D|H)}{P(D)} \tag{2.18}$$

One needs to know the probability that, if the hypothesis is true, what are the chances of observing that certain data? This is represented as the likelihood $P(D|H)$ as shown in the above formula. We also need to know the prior probability, that is what were the chances of the hypothesis being true before we observed the data (the prior probability), and the probability of observing that certain data regardless of whether the hypothesis is true or not which is known as the marginal likelihood (Prieto & Whittaker, 2013). The marginal likelihood, $P(D)$, can then be viewed as:

$$P(D) = P(H)*P(D|H) + P(\sim H)*P(D|\sim H), \quad (2.19)$$

where $P(\sim H)$ is the probability of the hypothesis not being true and $P(D|\sim H)$ is the probability of the given data occurring if the hypothesis is not true.

Looking back on the above formula, let us assume that a test for cancer is 80% accurate, that is, if one has cancer, the test is able to detect it 80% of the time. This would be the probability of getting certain data (a positive test) while the hypothesis is true (cancer is present) thus $P(D|H) = .80$ (Gelman et al., 2014). If we also have data that the percentage of people with this certain type of cancer is 2% then we know $P(H)$, the probability that, without any additional data, that one has this type of cancer. That leaves $P(D)$ which is the probability of getting a positive result on this cancer test regardless of whether one has cancer. If the chance of getting a false positive on the cancer test is 30% then we can calculate the probability of getting a positive on a cancer test by taking the proportion of having cancer (.02) times the proportion of a true positive (.8) and add it to the proportion of those not with cancer (.98) times the proportion of times you get a false positive (.3). This means that, regardless of whether you have cancer or not, you have around a 31% chance of getting a positive from the cancer test.

Filling in the formula we have:

$$P(H|D) = \frac{.02*.80}{.31} = .052 \quad (2.20)$$

That means that even if you did get a positive cancer test, you have only around a 5% chance of having cancer barring any other information. As can be seen, we start off with an original hypothesis (cancer is present), then alter the chances of that hypothesis being true by adding in additional data (results of a cancer test and the proportion of people who have that type of cancer). This example shows the general idea of the Bayesian approach. Expected outcomes are continually updated by recalculating the probability of a certain hypothesis given the past data when new data become available. This process does this by calculating the chances of multiple hypotheses with the updated data then seeing which hypothesis is the most probable with the current data. With new data being incorporated into prior probabilities to form an estimated posterior distribution, this posterior distribution then becomes the prior distribution for the next analysis conducted with newer data (Gelman et al., 2014).

While the example above used a single probability as a starting point for a prior distribution, the prior distribution often incorporates other information such as the variance of the distribution (Gelman et al., 2014). The prior mean and standard distribution for these parameters being estimated can be set either based on previous information or by allowing the standard deviation for the prior to be very high. A prior with a very high standard deviation is called an uninformed or noninformative prior because this acknowledges that the prior mean chosen may not be accurate and the large standard deviation for the prior allows for more fluctuation in finding the actual mean when the analysis is completed. This study will use these uninformed priors to try and ensure equivalency across the models.

As mentioned, the Bayesian approach has received criticism for the subjectivity that often involves priors. However, this subjectivity can be quantified by establishing priors relating to the

variance and mean of the prior distribution. If a prior is used that has strong research backing and evidence related to using that prior, then the prior distribution's variance would be smaller than if a prior was not as well supported (Lynch, 2007). The above Bayes' theorem remains largely the same but the singular probability is instead replaced with a distribution of probabilities. If Θ represents the prior distribution for a vector of parameters, then Bayes' theorem for computing the posterior distribution would be:

$$P(\Theta|D) = \frac{P(\Theta) * P(D|\Theta)}{P(D)}, \tag{2.21}$$

with *D* representing the dataset rather than a singular data point. What was previously known as the data likelihood, $P(D|\Theta)$, can now also be known as the sampling density of the data given the model parameters. $P(D)$ is a constant representing the marginal probability of the data distribution. Because it is a constant, the formula above can be simplified into showing how $P(\Theta|D)$, the posterior distribution, is proportional to the product of the prior distribution, $P(\Theta)$, and the likelihood, $P(D|\Theta)$

$$P(\Theta|D) \propto P(\Theta) * P(D|\Theta)), \tag{2.22}$$

where $\propto$ means "in proportion to". The formula can also be substituted with functions instead of probabilities which some Bayesian formulas mentioned later will use:

$$f(\Theta|D) \propto f(\Theta) * f(D|\Theta) \tag{2.23}$$

Basically, the posterior is proportional to the prior times the likelihood.

Calculating $P(D)$ is not as easy as the above formula would make it appear to be when distributions and data are complex as they often are. To approximate the posterior distributions, Bayesian sampling methods are used with one of the more popular ones being Markov Chain Monte Carlo (MCMC). A Monte Carlo simulation is a way to randomly draw samples based on a hypothesized distribution. A Markov chain is a method that allows for calculation of outcomes

based on a probabilistic series of events. Put together, the MCMC generates events that are dependent on each other based on probabilities making it suitable for Bayesian estimation (Brewer, 2009). Sampling from the posterior distributions of parameters (i.e., constructing a Markov chain) can be completed in a variety of ways.

One sampling method used in the data generation of a Markov chain is the Metropolis-Hastings algorithm. The MH algorithm draws a candidate value from a proposal distribution that can be easily sampled and decides to either accept or reject the value as the next value in the chain based on the acceptance probability. The acceptance probability is used to determine in which direction the distribution should go (Lynch, 2007). This algorithm works under situations where the posterior distribution is difficult to obtain through an analytical solution, and the proposal distribution is not symmetric. There are also other methods for sampling such as Gibbs sampling. This sampling can be used if there is a closed solution, meaning the distribution is easier to obtain. In Gibbs, the new value is always accepted (in other words the acceptance rate is 1) while the MH algorithm will either accept or reject the new value in the chain based on a weighted probability (Levy & Mislevy, 2016).

For this study Gibbs sampling was used. An example of a basic Gibbs sampling formula is explained below.

$$x = (x_1, x_2, \dots, x_k), \tag{2.24}$$

The above sets initial values for all parameters where *k* is the number of parameters. The first parameter is estimated as shown below:

$$x_1^j \sim p\left(x_1 \middle| x_2^{j-1}, x_3^{j-1}, \dots, x_k^{j-1}\right), \tag{2.25}$$

where *j* is the current iteration, and *j*-1 is referring to the values before the current iteration. The new value for parameter 1 for the current iteration *j* is calculated by drawing from

a distribution of parameter 1 values conditional on other parameters' values at the iteration *j*-1. Then the next parameter is estimated:

$$x_2^j \sim p\left(x_2 \middle| x_1^j, x_3^{j-1}, \ldots, x_{k-1}^{j-1}\right), \tag{2.26}$$

When the next parameter is being estimated, the value for all the previous parameters will either be the value estimated in *j* iteration, or if that value has not been estimated yet the value for that parameter will be set to *j*-1 iteration. This process continues until every parameter has been estimated.

With the new values, the process repeats where *j* is increased by 1 and then parameters are estimated again starting with the first parameter in formula 2.25. Then the formula estimates new values for each parameter and the process is repeated until parameters converge. Once convergence is reached, additional samples are drawn for use in point estimation of the parameters (Gearhart & Kasturiratna, 2018).

The MCMC method uses an iterative approach as well, and requires convergence of the analysis; that is, the posterior distribution needs to reach a stationary distribution. When the sampling method has completed enough sampling to represent the desired posterior distribution, then it is said that the process has converged. Assessing when that convergence happens is needed. This can be as simple as looking at the trace plot, a visual representation of the estimated parameter over the many samples and seeing when the parameter estimations become stable. An example of a trace plot is shown in Figure 2. A trace plot is also often used to determine burn-in sample size. A burn-in is a set of iterations done as a starting point that are then discarded before calculations are completed on all the other more stable iterations (Lynch, 2008).

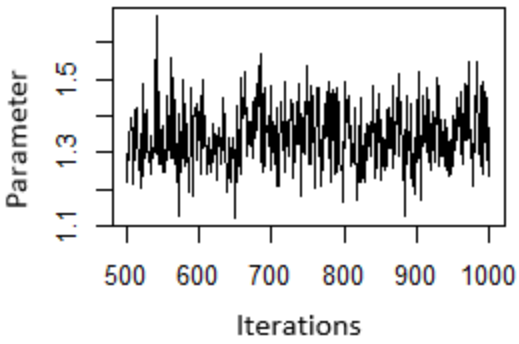

Figure 2. A trace plot for a parameter value with 500 iterations being "burned in."

More objective ways to assess convergence include analyzing the within chain variability of a single chain of iterations to the total variability when multiple chains are run. The idea here is that if the within chain variability takes up a large portion of the total variability, it means that the multiple different chains are converging to the same point and thus convergence is met. If you were to calculate within chain variability as *W* and the between chain variability as *B* with *T* being the number of iterations, then you can calculate a number to signify if the sampling is converging by:

$$\hat{R} = \sqrt{\frac{((T-1)/T)W + (1/T)B}{W}}, \tag{2.27}$$

where $\hat{R}$ is a "reduction factor", also called the Gelman-Rubin convergence diagnostic. If $\hat{R}$ is closer to one then the chains were able to converge to the desired posterior distribution (Levy & Mislevy, 2016). A boundary of $\hat{R}$ being less than 1.1 is often used to determine if convergence has been met (Brooks & Gelman, 1998). For this study, the model was set to auto-converge until the suggested boundary of 1.1 was met.

Finally, after the data have been generated and a posterior distribution has been determined, Bayesian estimation procedures use that distribution to make inferences about the population distribution simply by calculating the point estimates of the posterior distribution.

The parameters estimated from the Bayesian estimation procedures reflect the probabilistic properties of the posterior distribution (Levy & Mislevy, 2016).

One procedure to assist in evaluating the model fit is the posterior predictive model checking (PPMC) procedure. This uses the model derived from the observed variables to generate new data. This new data is then compared to the data the model was estimated from. If the model does well the discrepancy function based on the new data should be similar to the discrepancy function based on the original data. There are many ways to define the discrepancy function. A posterior predictive p-value is computed, and if significant (e.g., p-value < .05) it means the simulated data does not fit the original data and the model is inadequate (Gelman, Meng, & Stern, 2010).

Model selection methods can also be used when looking to see how model fit compares across models. Common model selection models that can be used for Bayesian analysis include information criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC), deviance information criterion (DIC), and Watanabe-Akaike information criterion (WAIC) (Gelman et al., 2014). These information criteria are useful for comparing how well certain models work for a dataset while adjusting for overfitting. However, since this research was focused on comparing the usefulness of a variety of models, these fit indices have limited usefulness for this study as some statistical models that are being investigated do not have fit indices.

Bayesian IRT formulation. Bayesian IRT incorporates priors when estimating (Fox, 2010). A Bayesian IRT 2PL model assigns prior distributions to the two parameters in the model. Item difficulty tends to be normally distributed (*N*) thus that the prior is:

$$b_j \sim N(\mu_b, \sigma_b^2)\,, \tag{2.29}$$

where *j* represents a specific item. The discrimination parameter prior can be defined by a log-normal distribution (Patz & Junker, 1999). This is symbolized by:

$$a_j \sim lnorm(\mu_a, \sigma_a^2) \tag{2.30}$$

The prior for the ability (or theta) distribution can also be set:

$$\theta_i \sim N(\mu_\theta, \sigma_\theta)\,, \tag{2.31}$$

where theta (θ) is the ability estimate for the individual *i*. The mean and standard deviation of the whole theta is represented by μ_θ and σ_θ. The likelihood of getting a certain response for an item is defined by:

$$p\left(x_{ij} \mid \theta_i\,, \omega_j\right)\,, \tag{2.32}$$

where x_{ij} is the response for participant *i* for item *j*. The parameters for the individual, *i*, is represented by θ_i and the parameters for a certain item *j* is represented by ω_j where ω_j would be b_j for 1 PL or a_j and b_j for a 2 PL model. One could also set all priors for all items to be the same. Using the previous definition of likelihood for this case, the likelihood can be thought of as the probability of an individual having a certain outcome on an item given an ability level and item parameters and priors.

Putting all the parts together you would end up with the posterior distribution proportionate to the prior distribution times the likelihood:

$$p(\theta, b, a \mid x) \propto \prod_{i=1}^{n} \prod_{j=1}^{J} p(\theta_i) p(b_j) p(a_j) \times p\left(x_{ij} \middle| \theta_i, b_j, a_j\right) \tag{2.33}$$

Essentially, the above calculates the posterior distribution by multiplying the priors times the likelihood for all individuals (*n*) and for all items (*J*) (Levy & Mislevy, 2016).

The theoretical interpretation for frequentist and Bayesian statistics differs (Lynch, 2007). Bayesian statistics seek to estimate the probability distribution rather than a single point

probability that frequentists' statistics estimate. Bayesian statistics view parameters relating to the data as random while the data itself is static, whereas frequentists view the data as a random sample from a true distribution while the parameters that the data are trying to estimate remain static (Lynch, 2007). This means that confidence intervals for Bayesian statistics, termed as credible intervals, view the range calculated as containing the parameter being estimated within a certain probability while frequentists interpret the confidence interval as the likelihood of obtaining the parameters from repeated sampling. While frequentists view the parameter distribution as having standard errors, Bayesian statistics view them as having standard deviations. Mathematically, these two interval measures are similar but their interpretation differs depending on whether the analysis was Bayesian or frequentist (Levy & Mislevy, 2016).

Bayesian models application. The use of prior information in Bayesian formulas can be either an advantage or disadvantage over more traditional frequentists' models. The use of reliable prior information can help give better estimations than frequentists' models may provide, even with a smaller sample size (Lee, 2007). While accurate prior information can help aid in the accuracy and speed of estimating parameters, inaccurate prior information may lead to less accurate predictions (Depaoli, 2014). Also as mentioned, Bayesian approaches do not have assumptions about the data being normally distributed and may help with datasets that do not meet the assumptions frequentist analysis requires (Gelman et al., 2014)**.**

While the Glicko approach mentioned in the Elo section could be considered a sort of Bayesian upgrade to Elo as it changes a previous static estimation of a parameter to one that changes as new data are gained about the player (Glickman & Jones, 1999), Glicko would be considered an analytical Bayesian inference as it does not use approximation methods such as MCMC, and therefore social scientists may not consider it true Bayesian. Instead, using

Bayesian formulas to predict skill ratings of players may look something like this (Coulom, 2008):

$$p(\gamma|G) = \frac{p(\gamma) * P(G|\gamma)}{P(G)}, \tag{2.34}$$

where $P(G|\gamma)$ represents the likelihood, in this case the probability of a win, and $p(\gamma|G)$ is a posterior distribution of γ (player's ratings) and $p(\gamma)$ is the prior distribution (Coulom, 2008). $P(G)$ is designed to function similarly to the marginal likelihood in the Bayes' theorem and acts as a constant. Essentially, every match the player completes is an additional data point where Bayesian estimation procedures use an iterative process that continually updates priors and updates the outcomes of these formulas. There are also other formulas that use Bayesian approaches to build on the Elo formula such as TrueSkill and Whole-history rating (Coulom, 2008; Dangauthier et. al., 2007).

The TrueSkill rating system uses a Bayesian approach in order to predict ability. It also adds in additional parameters. As other models mentioned in this study, the TrueSkill model has priors that are not only based on the mean estimate for a person's skill, but also on the variance or certainty of that skill estimate similar to what Glicko does. An additional parameter based on how much of the game outcome is due to luck is also in this formula (Dangauthier et. al., 2007; Herbrich, Minka, & Graepel, 2007).

Another more advanced gaming statistical model is the Whole-history rating system. This model works similarly however it allows for the system to go back and change participants' skill ratings even without additional game information for said participant. The logic being that since skill ratings are based on paired comparisons, if more information about one of the pair is collected, the formula should go back and update the other pair's information (Coulom, 2008). TrueSkill Through Time (TTT) does something similar (Dangauthier, Herbrich, Minka, &

Graepel, 2008). While the TrueSkill models and Whole-history estimations are gaining popularity in the gaming fields, this study is focused on the simpler gaming statistical models. Future studies may want to look at these more advanced gaming models.

A Comparison of Elo, Glicko, IRT, and Bayesian IRT Models

There has been research on comparing Elo, IRT, and Bayesian estimation procedures and their prediction results (Antal, 2016; Veldkamp & Matteucci, 2013; Wauters et al., 2011; Wauters et al., 2012). While IRT, and by extension Bayesian IRT, are generally accepted as more accurate measures of skill than Elo and tend to be most used in the social science fields like education, there has been research into how the social science fields can benefit from utilizing Elo related models. For instance, one study compared an IRT inspired Elo model, IRT, and proportion correct systems for testing purposes (Antal, 2016). The researcher viewed the simpler model of Elo as a possible alternative for situations where the extensive testing and sample size needed for IRT might not be viable. Elo tended to be highly correlated with IRT ratings (.942). However, it took Elo a larger number of matches than IRT to get to a stable skill rating. Additionally, the research had a small sample size of 137 students. While the article did have a small simulation part, it did not compare proportion correct. The real data analysis did include all statistical models of interest, but as mentioned, using just the real data may be a limitation as there is no known true ability and the correlations between the estimations makes assumptions about which measure is the most accurate across all conditions (Antal, 2016).

Wauters et al., (2012) also compared Elo and IRT models using a bootstrap study. Again, Elo correlated highly with IRT estimations varying from .85 to .90. The value of the correlation depended heavily on the weighting used in the Elo calculations, however. The identification of the weighting values was made by testing four weighting options in the Elo formula then

choosing the option with the highest correlation to true ability level. Additionally, this article managed to get IRT estimations to run on incredibly small sample sizes and even reported IRT estimates for a sample of 20 with a test of 25 items (Wauters et al., 2012). This indicates that either the article may indicate possible errors or that IRT can converge using very small samples making the use of Elo for those types of data less appealing.

A positive view of the above studies might lead us to conclude that Elo models may produce similar results to IRT models under certain conditions. However, the above studies have a limitation in that IRT is used as the comparison with the assumption of it being the most accurate model in all conditions, even in conditions with smaller sample sizes, rather than comparing all model estimates to known ability values. Even with studies that used bootstrapping on a large dataset, the comparison of Elo and proportion correct were correlated with IRT parameters that were calculated on the complete dataset (Wauters et al., 2012).

There have been few simulation studies looking at Elo and IRT models that compare their estimates to known ability values (Pelánek, 2014). One study found that when using simulated data comparing Elo to the Rasch model, the Elo estimates were noticeably worse than the Rasch model estimates. However, when a variation of the Elo formula was used that replaced the constant K with a parameter that included uncertainty, the correlation between the estimates of the two models were largely similar with correlations being mostly above .99. This study also explored how Elo was similar to Bayesian knowledge tracing and overall this study found Rasch, Bayesian knowledge tracing, and Elo to be largely similar (Pelánek, 2014). This study also looked at proportion correct and found that while IRT and Elo were both highly correlated with the true ability, proportion correct tended to be about .10 less of a correlation to true ability across sample sizes. If Elo could outperform proportion correct as mentioned in the one study

(Pelánek, 2014), Elo could serve as a possible ability estimation model over classical test theory estimation in the event that the data may not be suited for IRT estimation.

Another study that had a small simulation aspect also had similar findings, where IRT and a modified Elo resulted in similar estimates even though it required a higher number of questions for Elo to get to those estimations (Antal, 2016). Antal recommended 25 items being the amount of data needed for reliable Elo estimates. This may lead to an interesting property of Elo where its strength can be that it can work with small sample sizes but needs a larger set of items or matches to get close in accuracy to IRT. Since IRT uses iterations and more complex estimation methods, if the sample size is too small, estimates may not converge where the simplicity of Elo allows for estimation of abilities even if they may be less accurate.

Another study compared IRT results with the results of ability estimation parameters obtained through various Elo formulas. One formula was the Elo formula that Brinkhuis and Maris (2009) used which allowed for a logistic function to be used in the formula for weighting rather than have the weight be a static parameter. The Wauters (2011) study compared the differing weights that could be used to see how they correlated with the IRT parameters. This study changed the base used in the Elo model to more closely resemble the Rasch model meaning the expected probability of winning, or E, would be calculated as:

$$E = \frac{1}{1 + e^{-1(P-O)}} \tag{2.35}$$

Thinking back to the player rating update formula:

$$Post = Pre + K(Out - E)\,, \tag{2.36}$$

K would no longer be a static number but would be calculated by:

$$K = \frac{K_o}{1 + a \times e^{(b*N_{ip})'}}\,, \tag{2.37}$$

where *a* and *b* are parameters similar to IRT parameters with *a* and *b* parameters related to the properties of how quickly estimates of ability would converge. N_{ip} is the number of matches or items completed before the participants' current match/item. The $'$ indicates a derivative. Overall, there were high correlations between the Elo variation formula and IRT estimations ranging from .80 to .94. This is additional evidence that an Elo type of formula can be useful in predicting ability in instances where IRT might not be feasible. However this study is also cautionary as simply changing the weighting parameters in the Elo formula allowed for a variety of estimation parameters, some far less accurate than others (Wauters et al., 2011).

There has also been exploration in combining and comparing Elo and Bayesian approaches. One study looked at how the use of a Reference Agent Space in chess games can allow for a system that does not have the limitations that Elo has. It can even detect differences in skill level within a single game. However, their methods rely on a set definition of what good and bad moves in a game can be (Fatta, Haworth, & Regan, 2009). This may be a relatively simple thing to determine in chess where there are specific moves one can make that can either negatively or positively affect the outcome of the game, but in situations with a more open environment it would be very difficult to utilize this model, if even possible. Still, this study is just one example of research exploring how to improve traditional game match making systems (Fatta, Haworth, & Regan, 2009).

There have also been studies looking at IRT and Bayesian IRT. One study advocated that using prior information in addition to IRT with computer adaptive testing (CAT) could save time and money by lowering the number of items needed to obtain an ability estimation with a small standard error (Veldkamp & Matteucci, 2013). As mentioned in the article however, there are some ethical implications on using information from the participant that is not solely from the

data gained from the test. One study that looked at how Bayesian IRT and non-Bayesian IRT compares found that Bayesian estimation procedures produced more accurate estimations (Gao & Chen, 2005). However, it is worth noting that using Bayesian estimation procedures has an advantage over IRT with non-Bayesian estimation procedures, beyond just possibly better accuracy, in that prior information can help with achieving convergence in situations where non-Bayesian IRT may not converge. Use of these different models may depend on the situation rather than which model is more accurate with ideal data.

While the most extensive and complicated statistical models may end up with the most accurate estimates, there are situations where using those models may not be ideal. More complicated statistical models commonly require larger samples and numbers of responses to accurately obtain their estimates. This is why the research above comparing the different statistical models often looked at other outcomes such as speed of calculations and how number of responses relates to accuracy. As mentioned, even though it may take more items/matches for Elo to get a rating as accurate as IRT, Elo can provide estimations for small sample sizes while IRT methods may not mathematically converge to provide estimations. Additionally, many of the models are conceptually similar to each other with the only differences being having certain parameters in the model fixed or not. In particular the Elo and IRT 1PL mathematical models are very similar but IRT models generate estimations based on a set of items while gaming models like Elo make estimations after every match. Finally, the restriction of model assumptions are a factor in determining the appropriateness of models under certain situations.

Although there has been research comparing IRT, Bayesian IRT, and Elo models, it has not been extensive and often is focused on a single outcome such as accuracy in comparison to the assumed “best” estimate since it is conducted on empirical data where the true ability is

unknown (Antal, 2016; Wauters et al., 2011; Wauters et al., 2012). Using simulation data in order to know the "true" ability may be one way to compare these models without the assumption that one estimation will always be more accurate than the others. There does not seem to be many simulation studies looking at gaming statistical models for analyzing achievement or psychological types of data in comparison to more traditional social science psychometric models (Antal, 2016; Pelánek, 2014). One weakness of simulation studies is that simulation of the data partly relies on which formula you use to predict outcomes which in turn is related to the statistical models you are comparing. This may lead to some variation in findings as an empirical data study found proportion correct to be a better choice than Elo (Antal, 2016; Wauters et al., 2012) while a simulation study found Elo to be a better choice than proportion correct (Pelánek, 2014). This shows some need for using both simulation and empirical data to complement each method of analysis' weaknesses and to better identify trends.

Another gap in the research has been a lack of focus on how educational data and gaming data may interact with different statistical models. While educational datasets may benefit from simpler gaming statistical models due to limitations such as sample sizes, gaming datasets may benefit from the use of more complex social science statistics procedures. As mentioned, many popular and commonly used gaming ability estimation models originated in the 1970s with the assumption that "easy to compute" calculations were essential to easily updating tournament data (Glickman, 1995). With access to better computational power and many gaming datasets having very large sample sizes, it would be interesting to see how more complex statistical models, such as IRT and Bayesian models, compare with Elo under various conditions.

This study was designed to compare a variety of models on accuracy and variability outcomes using both real and simulated data in order to synthesize research being completed

across a variety of fields and provide suggestions as to which models may be the best choice depending on the situation. Simulations were focused on comparing the correlations between the model estimates and the true ability. The real data analyses were an investigation of the correlations between the ability estimates from the different models, with a sample of the real data being set aside to compare how well the models predict the outcomes of that subset. This research hopes to address some of the short-comings of the literature and to demonstrate how the statistical models selected for this study compare to one another under different conditions in order to show the potential benefits of each model.

CHAPTER 3

METHODS

The purpose of this study was to compare a variety of outcomes using ability estimation models from social science and gaming fields. This study used both real data and simulated data with binary outcomes, such as winning or losing a game or getting an answer right or wrong, to investigate the effectiveness of seven models in estimating ability or skill level. The models being compared are proportion correct/win, basic Elo, Glicko, IRT 1PL/2PL, and Bayesian IRT 1PL/2PL.

Simulation Study

The conditions that vary for the simulation study are data generation techniques, sample size, and match/item size. The data generation generated Elo, Glicko, IRT 1PL, or IRT 2PL data. The sample sizes varied between 50 and 150 for all data generation techniques but also had a 500 sample size for the IRT 1PL and 2PL generation techniques. The match/item sizes were 5, 15, and 30. There were a total of 30 conditions.

Sample Size

Data for 50 and 150 individuals were generated. A sample of 50 would be common in a smaller gaming tournament or the number of students a teacher might have for a specific area of study (e.g., students in pre-algebra). A sample size of 50 would be a small sample for more data intensive analysis such as IRT and Bayesian IRT but it might be sufficient for more simple analyses like the gaming estimations which are often used in tournaments with smaller sample sizes (Flateby, 1996, Sinharay, Johnson, & Stern, 2006; *The ACT technical manual*, 2017; "Upcoming Tournaments," n.d.). Samples of 150 are likely for smaller research studies, students within a school taking a specific course, or mid-size tournaments for games. One hundred and

fifty would be a sample size that may be sufficient, but still small, for more of the advanced IRT models (Chang & Davison, 1992; Finch, 2011; "Upcoming Tournaments," n.d.). To demonstrate a trend for the educational data, sample sizes of 500 were also included for the IRT 1PL and 2PL data generations.

Item/Match Characteristics

Educational testing and gaming match data were simulated for three test lengths and number of matches (5, 15, 30) focusing on assessment of abilities with small numbers of items or matches to moderately larger tests or tournaments (e.g., ACT, 2007; Ansley & Forsyth, 1985; Finch, 2011; Sinharay, Johnson, & Stern, 2006; "Upcoming Tournaments," n.d.). Item difficulty for the education-type data was generated using a normal distribution (~N(0,1)) with discrimination values set at 1 (e.g., Bolt & Gierl, 2006). For the 2PL education data, discrimination values were allowed to vary using a log-normal distribution [~lnorm(0, .5)] similar to other IRT simulation studies (Miller & Oshima, 1992; Reckase & McKinley, 1991).

The gaming data included the same number of matches (5, 15, 30) as that generated for the educational testing scenario. These numbers were chosen because it would be important for games to be able to match players after only a short number of matches (e.g., 5) in order to keep the player's interest, while 15 and 30 matches can be typical of both tournaments and of short to moderate cognitive subtest lengths (Ansley & Forsyth, 1985; Flateby, 1996). Fifteen matches/questions is a middle ground and past research has found that around 20 items can produce estimates of Elo and IRT that tend to be similar (Antal, 2016).

Data Simulation

For the educational testing situations, a unidimensional set of initial ability estimates (~N(0,1)) were used as the participants' latent ability skill level for creating a string of responses

to sets of items previously described. The true ability response sets were generated using the psych package in R version 3.5.1 for the IRT 1PL and 2PL data (Revelle, 2018).

For the estimation of player ability in the gaming situations, players were randomly matched with each other with the player outcome depending on the logistic probability of winning based on the latent ability assigned previously using a ~N(0,1) distribution. That is, if player 1 had a 2% chance of winning then the player outcome was randomly drawn from a distribution that represented that chance using the basic Elo Formula for probability of outcomes. Glicko was simulated similarly but the percent chance of a win also included a player rating deviation parameter and not just mean difference between the ratings. For the constants present in the Elo and Glicko formula when generating and estimating the data, the default parameters were used (Glickman, 1995).

The gaming data included the same number of matches (5, 15, 30) as that generated for the educational testing scenario. However, the gaming data are different than educational data in that players are not matched with every other player while in the educational data every participant answered every item. This means that the gaming matrix of matches is complex and sparse with each player matched with different combinations of other players. As such, it was currently not realistic for coding purposes for every person to have exactly 5, 15, or 30 matches due to how the players' data were intertwined. Therefore, the data were generated in a way where the average number of matches the players had was either 5, 15, or 30 with some players having more matches and some having fewer.

There was one adjustment made in the gaming data to allow for IRT 1PL and 2PL estimates to be calculated. Due to the way IRT abilities are estimated using the TAM package in R, two dummy players were added to the simulated gaming dataset with one losing all matches

and one winning all matches. When IRT is applied to gaming data, players are considered both participants and "items." When estimating one player's ability, all other players are considered the "items" that they are matched against. Similarly, when other players' abilities are being estimated, the former player is now the item. This results in an n x n mirrored inverse matrix. However, the IRT package does not provide an estimate when everyone gets an item wrong. Since players are also items and the number of matches in this simulation could be as low as 5, it is likely to have people who won 0 matches which the package would read as an item everyone answered wrong. Therefore, two players were added with one winning all matches and one losing all matches against all "items" to allow for ability estimates for all other players. The two dummy players' abilities were not included in the simulation results. In the educational data, it is unlikely to have an item that all people get incorrect even with the smallest sample size (n = 50) and largest number of items (i = 30) condition used in this study. In the event that the educational data contained an item where everyone answered it wrong, that iteration was skipped.

There are a couple of factors to consider in the comparison of the ability estimates using the different models with the gaming and educational data. While the gaming estimates update ability ratings for participants after every game, IRT and Bayesian IRT estimate ability using a complete set of items. As a result, item or match order may show an effect on the estimates in gaming statistical models (e.g., Elo, Glicko) while it should not when using IRT and Bayesian IRT though order effects with the gaming statistical models should be small. It is also worth remembering that the gaming and educational data have different matrix structures. The educational data has participants with complete data on the test questions, while for the gaming data not every participant was paired with every other participant. This led to the gaming data

being sparsely distributed within a population matrix of response strings, while the educational data was a matrix of complete response strings.

Simulated Data Analysis

The skill rating estimations from seven gaming and education models (Elo, Glicko, proportion correct, IRT 1PL/2PL, Bayesian IRT 1PL/2PL) were compared with the true ability ratings. Correlations between true and estimated abilities were provided. The TAM package for R version 3.5.1 was used for conducting the IRT 1PL and 2PL analyses and maximum weighted likelihood estimation was used (Robitzsch, Keifer, & Wu, 2018). JAGS and an R package that runs JAGS (rjags) was used to run Bayesian IRT 1PL and 2PL estimations and Gibbs sampling was used (Plummer, Stukalov, & Denwood, 2018). Uninformed priors were used for Bayesian IRT 1PL and 2PL for both simulated and empirical data analysis, ~N(0, .01 [precision]). This was to test if the basic Bayesian IRT could improve upon other models without certain known priors. The PlayerRatings package in R was used for Elo and Glicko estimations (Stephenson & Sonas, 2016). There were 500 replications for each condition. For the Bayesian IRT models, all models were run to auto-converge according to the convergence criteria of 1.1 (Brooks & Gelman, 1998). There were 1000 iterations used for burn-in. An overview of the simulation conditions is provided in Table 1.

Table 1

Summary of Simulation Conditions

Variables	Conditions
Data Generations Procedures	Educational Data (IRT 1PL and IRT 2PL);
	Gaming Data (Elo and Glicko)
Sample Size	50 150 500*
Match Size	5 15 30
Estimation Analysis Models	
Gaming Statistical Models	Social Science Statistical Models
Elo	Proportion Correct
Glicko	IRT 1PL
	IRT 2PL
	Bayesian IRT 1PL
	Bayesian IRT 2PL

* Samples of 500 were only conducted for educational data comparisons as gaming matrices with 500 matches had too large of a time requirement even when using a high-performance computer.

Real Data Analysis

Educational Testing Data

The educational testing data set used in this study was the mathematics data from the 2011 Trends in International Mathematics and Science Study ([TIMSS], International Association for the Evaluation of Educational Achievement, 2011). TIMSS is a collection of cognitive, attitudinal, and background data of 4th and 8th graders from a variety of countries. One booklet administered to a subgroup of the US participants in the 8th grade was used so that a common set of items could be investigated. A booklet with mathematics items was selected. The

TIMSS mathematics test has four content areas for items: numerical, algebra, geometry, and data/chance. Numbers of items that matched the simulation conditions were used for estimation (5, 15, and 30), and 3 items that represent varying difficulty levels were used to identify levels of accuracy in correctly predicting a correct answer.

The models were used to estimate item parameters for all 33 items. Proportion correct was used to select an item at medium difficulty (difficulty of or around .5), hard difficulty (.25) and easy difficulty (.75). These three items were set aside, along with their item parameter estimates to be used in the prediction analysis. For the prediction analyses, 5 items were used for the 5 item sample, then 15, and then 30 items for estimating participants' thetas. The 5, 15, and 30 items were randomly chosen but in a way that each subset would have a composition of items similar to the full 33 questions in regard to content areas. The thetas were then used to calculate each participant's chance of correctly answering the 3 items set aside in order to calculate prediction accuracy using the 3 samples of 5, 15, and 30 items. The item parameters for the set aside items were the parameters estimated using all 33 items and the model used matched the model being analyzed. When predicting the outcome on the set aside items, however, the probability of a correct answer for the formula used was always the IRT 1PL formula. This is primarily due to real gaming data analysis using values in the dataset as its "true value" and since the dataset only gives ability parameters this means formulas with discrimination values (2PL) or rating deviations (Glicko) cannot be used to calculate probability of a win. Kappa coefficients were calculated to identify the proportion of correct predictions made for the three items beyond random chance. The kappa coefficient is often used for comparing classification outcomes on multiple test forms (criterion-related reliability) or interrater agreement on dichotomous outcomes. When interpreting the kappa, the values are considered a minimal effect being from

.01 to .2 and a weak to moderate value being from .21 to .4 (Viera & Garrett, 2005). This process was done twice with one subset of three items being algebra and the other subset being from the numerical category in order to better see trends.

For using gaming estimation procedures with the educational data it was similar. Elo and Glicko were used to calculate the "players'" abilities, which were actually items being treated as players, using the complete data and these were used as the item parameters after being standardized. The same three items were set aside, and the players' abilities were used to predict the outcome on the 3 items of varying difficulty using the 5, 15, and 30 item response sets. The IRT 1PL formula was still used to calculate the probability of a correct answer for these set aside items. Kappa coefficients were again used to calculate prediction accuracy for the different conditions.

Gaming Data

The empirical gaming data set used for the study is a collection of over 20,000 chess games from an online chess game site called Lichess. The data were collected using open source API data collection and posted on Kaggle, a dataset sharing website (J, 2017). Skill ratings were assigned to the players using four models (Elo, Glicko, proportion correct, and IRT 1PL). As will be discussed later, the larger dataset made it difficult to run the 2PL models and the Bayesian 1PL model. These skill ratings were then used to predict outcomes for the player's next match. The Elo formula was used to predict the probability of a win for the next match by using the ability estimates from either the Elo, Glicko, proportion correct, or IRT 1PL models. The models made estimations after a certain number of matches (5, 15, 30). The outcome for prediction were the outcomes of the 6^{th}, 16^{th}, or 31^{st} game. Only players who completed 31 unique games were

used for calculating kappa coefficients. The accuracy of the predictions of the different models were compared between the models' predicted outcomes (win/loss) and actual outcome.

For more specifics in regard to estimations with gaming data, similar to the process used with the IRT estimates, the most accurate ability for the opponent needed to be decided on. When predicting player one's gaming outcome at match 6, 16, and 31, player 2 had their ability estimate assigned using the skill value estimated for them in the original dataset of 20,000 matches while player one's estimates (the players with at least 31 matches) were their ability after 5, 15, or 30 games. This is due to the number of matches the players have in the dataset being very positively skewed making it likely that the opponent at the 6^{th}, 16^{th}, or 31^{st} match may only have a few games in the dataset. Since the dataset did not have all matches played by every player, a decision was made to use the player rating given in the dataset which included other matches the player had completed prior to collection of this dataset. These player ratings were used for the "opponents" because they were deemed the more stable and most appropriate estimate to compare to. The Lichess site used the Glicko 2 skill system and this estimate was rescaled for comparison to the other estimations.

In gaming data, there is a factor called "first move advantage" that exists which is based on whether one is assigned to white or black chess pieces. Players were randomly assigned in the real dataset as either black or white, therefore "first move advantage" should not be an issue in the current study (Micklich, 2009).

Similar to the simulated gaming data, the real gaming data included participants that have not all played against each other, therefore the matrix was sparse. This means that unlike the educational data, three players of differing ability levels cannot be set aside since it is unlikely to

find 3 players who have faced all the other players. Therefore, the next players' match was evaluated when predicting match outcome after 5, 15, and 30 games.

Supplemental Data Analyses

The amount of time for calculating estimates was evaluated as a practical measure of efficiency that is of high interest to the gaming community and likely of secondary interest to the educational community. This information should supplement data analyses of accuracy though it is important to keep in mind that length of time to run analysis depends on the computer and program used and not just the statistical model.

CHAPTER 4

RESULTS

Starting with the simulation study, the overall trend of how model estimates correlate with true ability are provided by aggregating over study conditions. These trends are compared in relation to how they varied across different data generation conditions. Variations in these trends based on sample size and number of items/matches will then be presented. After simulation data results are provided, the results for the two types of empirical data comparisons will be made across the models and item/match sizes.

Simulation Study Results

Data are first provided by aggregating over sample size conditions of 50 and 150 and item/match sizes of 5, 15, and 30 for the different ability estimation models (Elo, Glicko, proportion correct, IRT 1PL/2PL, and Bayesian IRT 1PL/2PL). These results are separated by data generation methods (Elo, Glicko, IRT 1PL, and IRT 2PL). Table 2 provides a summary of the average correlations between true and estimated ability across the six conditions for each of the statistical models differentiated by the data generation methods used.

Table 2

Correlations between True and Estimated Ability Averaged across All Sample Size and Item or Match Conditions

	Data Generation Procedure			
Ability Estimation Model	Elo	Glicko	IRT 1PL	IRT 2PL
Elo	.793	.793	.810	.804
Glicko	.731	.731	.689	.671
Proportion Correct	.785	.785	.818	.812
IRT 1PL	.779	.779	.815	.809
IRT 2PL	.670	.670	.783	.787
Bayesian 1PL	.801*	.801*	.819*	.813*
Bayesian 2PL	.652	.652	.726	.768

Notes. Highest correlation for each condition is indicated by a *.

Results indicate that Bayesian IRT 1PL estimates tended to be the most correlated with true ability across data types, while Glicko tends to be on average poorer across data types but especially with educational data. Elo, proportion correct, IRT 1PL and Bayesian IRT 1PL tended to have similar correlations between estimates and true ability when averaged over all conditions. Elo tended to outperform IRT 1PL with the gaming generated data but IRT 1PL tended to outperform Elo with the educational data. The educational data overall displayed higher estimated and true ability correlations than the gaming generated data. The Elo and Glicko generated data resulted in nearly identical correlations with differences between the correlations showing only when taken to around 7 decimals. Due to this, in further tables, when the Elo and Glicko resulted in identical estimations when using 3 decimals, the Glicko results will be absent and the Elo results with be marked with an asterisk. When all of the correlations are aggregated,

it appears the 2PL model estimates have the lowest relationship with true ability for the gaming type data.

Sample Size and Number of Items or Matches

In this section, comparisons are made between ability estimation models distinguishing between sample sizes and number of items or matches while aggregating across data generation method. Table 3 provides output for correlations between estimated abilities and true abilities for each of six sample and number of items/matches condition combinations, averaging across all four data generation methods. Similar to Table 2, only sample sizes of 50 and 150 are included, making this a comparison for smaller samples.

Table 3

Correlations between True and Estimated Ability Based on Number of Items or Matches

	Sample Size					
	50			150		
	Number of Items or Matches					
Model	5	15	30	5	15	30
Elo	.665	.840	.904	.654	.835	.901
Glicko	.649	.774	.789	.602	.705	.713
Prop. Correct	.654	.843	.913	.643	.839	.909
IRT 1PL	.637	.844	.915	.625	.839	.912
IRT 2PL	.498	.779	.884	.516	.788	.892
Bayesian 1PL	.668*	.850*	.917*	.656*	.846*	.915*
Bayesian 2PL	.508	.710	.827	.537	.755	.860

Notes. Highest correlation for each condition is indicated by a *.

For all statistical models, an increase in match or item size led to better correlations between estimated and true ability. While just 5 matches/items led to poor correlations across

models, the 15 match/item condition tended to have moderate correlations between true and estimated abilities with the single parameter models having a correlation mean near .85. The 30 match/item condition had the strongest correlations and the only conditions with correlations greater than .90. The models that appeared to be affected most by match/item size were the IRT 2PL and Bayesian IRT 2PL models. They tended to estimate poorly with small match/item sizes. However, with 30 matches/items, the 2PL estimates were closer to, but still less correlated than all models other than Glicko.

This pattern of the correlations between true and estimated ability is similar when looking at both sample sizes, that is, an increase in sample size did not increase correlations between estimated and true ability in most cases. However, impacts of sample size increases were different for the different models and data types. The increase in sample size seemed to have little effect, or may even show a decrease, on the correlations for all statistical models except the 2PL models where an increase in sample size led to an increase in correlations. Sample size differences by data generation method will be discussed in the sections where the academic and gaming data are investigated separately.

Looking at Table 3 with both sample sizes we can see that the highest correlations occurring overall were for the Bayesian IRT 1PL model. However, with the 30 item/match size condition, the IRT 1PL and Bayesian IRT 1PL were very similar with Elo and proportion correct being only slightly less than those correlations.

Figure 3 shows graphs of the model's correlations by data generation methods and match/item sizes, averaged across the sample sizes of 50 and 150, with the Glicko data generation excluded due to the correlations having the same results as Elo generated data. Trends for increase in true and estimated ability correlations are similar for Elo, IRT 1PL, Bayesian IRT

1PL, and proportion correct across the data generation methods as match/item size increases. Glicko estimations are best for gaming generated data, but tend to have lower correlations overall when compared with most other models. Further, Glicko does not seem to have the correlation with true ability increase as much with an increase in item/match size as was observed with other models. The 2PL models have low correlations for 5 match/item conditions and for data generated using the gaming models. The next sections differentiate results by model estimates, sample sizes, number of match/item sizes, and by type of data generated.

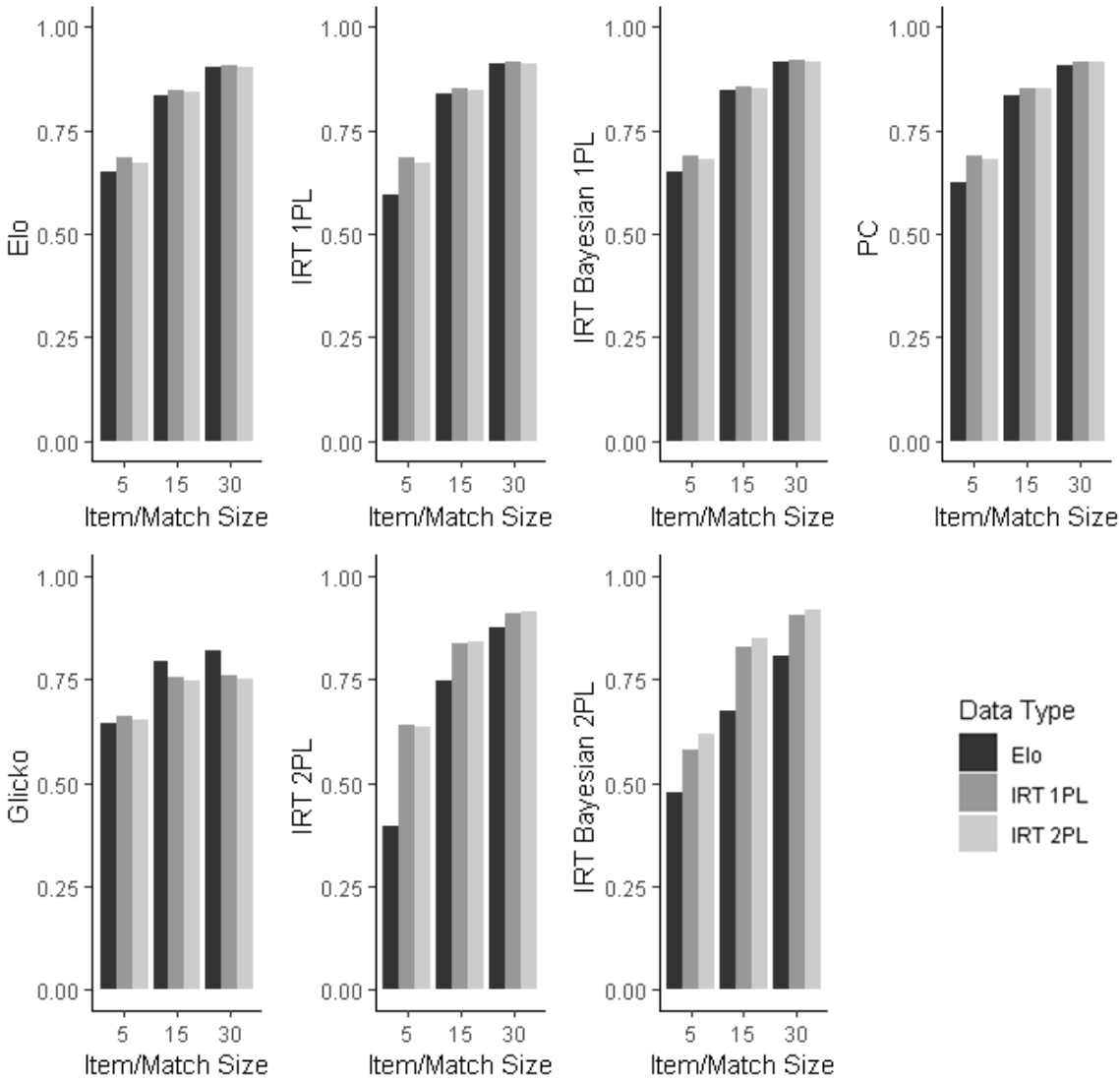

Figure 3. Correlations by data generation, match/item size aggregated by sample size

Elo and Glicko Gaming Data Generation

In Table 4, the results across the six sample and match study conditions are presented for the Elo and Glicko data generation procedures. Since the results for the Glicko data generation condition are extremely similar to that of the Elo data generation, only Elo is shown. The similarity between Elo and Glicko generated data is due to the Glicko procedure only including one additional parameter in the ability estimation process called the rating deviation which adjusts as additional matches occur. This deviation can be useful in ability estimation by beginning as a large value (e.g., 350; Glickman, 1995) for a participant's first estimate and then updating to a smaller value for each subsequent match with the idea being that as more matches occur, the rating given to the player is more certain. However, when simulating data, one is generating the outcome of a match based on two players' abilities, where the true ability is known. Including a large rating deviation in the estimation process appears counter-intuitive when true ability is known. Data were generated using a large deviation and then a smaller deviation of 1. When compared, using a large value of 350 resulted in data generation which led to all correlations between estimated and true ability being lower, regardless of the estimation procedure (Elo, Glicko, proportion correct, IRT 1PL/2PL, and Bayesian IRT 1PL/2PL). Thus, since true ability was known, the rating deviation was set to 1 (as small as possible) in order to generate data, with the value of 350 maintained for estimation procedures. More will be said about the usefulness of generating Glicko data in the Discussion.

Table 4

Correlations between True and Estimated Ability for Elo and Glicko Generated Data

	Data Generation Method					
	Elo*					
	50			150		
Model	5	15	30	5	15	30
Elo	.652*	.835	.903	.641*	.829	.900
Glicko	.642	.796	.821	.623	.743	.760
Prop. Correct	.625	.834	.910	.612	.828	.904
IRT 1PL	.595	.839	.914	.579	.833	.911
IRT 2PL	.395	.747	.877	.393	.737	.871
Bayesian 1PL	.652	.846*	.916*	.638	.842*	.914*
Bayesian 2PL	.478	.675	.803	.475	.672	.807

Notes. Elo and Glicko resulted in the same correlations up to 3 decimal places; highest correlation for each condition is indicated by a *.

When analyzing the gaming data, an increase in matches leads to increases in correlations across all statistical models. The range of correlations for the 5-match condition with the Elo generation procedure is .395 to .652 with 50 participants and .393 to .641 for 150 participants. When the number of matches increase to 30, the range of correlations increase from .803 to .916 for 50 participants, and .807 to .914 for 150 participants.

For the gaming data, the increase in sample size results in similar or slightly lower correlations between estimated and true abilities, even for the 2PL models. This decrease in correlations by sample size when controlling for match size is likely due to the format of the gaming data as an increase in sample size when controlling for match size leads to a sparser matrix. This will be discussed in more detail in the next chapter.

Overall, Elo tended to be one of the best estimations with the lower match size condition, but IRT 1PL and Bayesian IRT 1PL outperformed Elo at 15 and 30 match conditions. This aligns with the idea that the simpler gaming model may do better with very small match sizes while more complex models may produce better estimations with larger datasets with more matches. The Glicko did not provide accurate estimates of the true abilities for gaming data especially at higher match sizes when compared to most other models. The 2PL models also tended to do poorly with the gaming data and did worse than all the other models including Glicko.

While sample size seems to have little effect on correlations and may even result in a decrease in correlations, the standard deviation of the correlations from the 500 replications were affected by sample size as shown in Table 5. As sample size increased, the standard deviations of the correlations decreased making them more consistent even if the correlations did not increase overall. The effect of increased sample size reduced variability in correlations similarly to the effect of an increase in match size. The effect of sample size on the standard deviations of the correlations seemed to be consistent across statistical models.

Table 5

Standard Deviations of the Correlations for Elo and Glicko Generated Data

	Data Generation Method					
	Elo*					
	50			150		
Model	5	15	30	5	15	30
Elo	.070*	.040	.024	.041*	.023	.014
Glicko	.077	.052	.044	.046	.037	.036
Prop. Correct	.077	.039	.023	.050	.023	.013
IRT 1PL	.090	.040	.022	.052	.023	.013
IRT 2PL	.111	.058	.033	.059	.036	.021
Bayesian 1PL	.076	.038*	.022*	.045	.023*	.013*
Bayesian 2PL	.100	.083	.062	.057	.050	.036

Notes. Elo and Glicko resulted in the same correlations up to 3 decimal places; the lowest standard deviation for each condition is indicated by a *.

IRT 1PL and IRT 2PL Data Generation

When the data are generated to simulate 1PL and 2PL educational data, there are some trends that are similar to the gaming data but also some differences. Table 6 shows the comparison of IRT 1PL data across the three sample sizes and Table 7 shows the comparison of IRT 2PL data across sample sizes. Similar to the gaming data, as number of items increase in the IRT data, correlations between true and estimated ability increase. For the IRT 1PL data with 5 items, correlations are small, ranging from .517 to .689. As items increase to 15, correlations increase to .702 to .855. Relatively strong correlations (above .90) are observed for Elo, proportion correct, IRT 1PL and Bayesian IRT 1PL in the 30 item condition. Correlations above .90 are observed for IRT 2PL and Bayesian IRT 2PL, but only with the samples of 150 and 500.

Samples of 50 result in much smaller correlations for the 30 item condition for the 2PL models when compared to larger sample sizes. When comparing results for the 1PL and 2PL data generations, the correlations for both seem similar. One pattern is that the 1PL models give better estimations than the 2PL models on 1PL data across all conditions. However, with the 2PL data, the 2PL models begin to estimate better than the 1PL models with the larger number of items (15 and 30 items) and larger sample size conditions (150 and 500). The Bayesian IRT 2PL model had the best estimate for 5 items with a 500 sample size, but it was nearly identical to the Bayesian 1PL and proportion correct correlations. While Glicko tended to have lower estimations although similar at small item sizes, the increase in item size did not correspond to an increase in correlation to true ability estimation as much as the other models leading to Glicko performing poorly compared to the other models especially at larger item sizes.

Contrary to what would be expected, increasing sample size seemed to have little impact on most of the correlations. The 2PL models were the only ones where an increase in sample size led to a noticeable increase in correlations. While it makes sense that more complicated models like the 2PL models would benefit more from an increase in samples size, it was also expected that the other statistical models' correlations would improve with increased sample size. However, other research looking at the correlations between true and estimated abilities for IRT models with small samples and smaller number of items have found similar results and will be discussed in more detail in the Discussion chapter (Foley, 2010).

Table 6

Correlations between True and Estimated Ability for 1PL Generated Data

	Sample Size 50			Sample Size 150			Sample Size 500		
	Number of Items			Number of Items			Number of Items		
Model	5	15	30	5	15	30	5	15	30
Elo	.684	.847	.907	.674	.843	.904	.678	.843	.903
Glicko	.662	.756	.760	.612	.671	.674	.600	.653	.660
Prop. Correct	.689*	.854	.917	.681*	.852	.915	.685*	.852	.915
IRT 1PL	.685	.851	.916	.677	.849	.915	.681	.848	.915
IRT 2PL	.608	.806	.888	.642	.836	.909	.672	.844	.913
Bayesian 1PL	.689	.855*	.918*	.681	.853*	.917*	.685	.853*	.917*
Bayesian 2PL	.517	.702	.820	.578	.829	.908	.669	.850	.915

Notes. Highest correlation for each condition is indicated by a *.

Table 7

Correlations between True and Estimated Ability for 2PL Generated Data

	Sample Size 50			Sample Size 150			Sample Size 500		
	Number of Items			Number of Items			Number of Items		
Model	5	15	30	5	15	30	5	15	30
Elo	.673	.844	.904	.661	.838	.900	.666	.838	.899
Glicko	.651	.748	.753	.552	.663	.659	.588	.643	.643
Prop. Correct	.678	.851	.915	.669	.847	.912	.672	.848	.912
IRT 1PL	.674	.848	.914	.664	.842	.912	.667	.842	.911
IRT 2PL	.606	.819	.896	.635	.843	.915	.660	.853	.920
Bayesian 1PL	.678*	.852*	.916*	.669*	.848*	.913	.672	.848	.914
Bayesian 2PL	.557	.786	.880	.618	.848	.920*	.672*	.866*	.926*

Notes. Highest correlation for each condition is indicated by a *.

It should be noted that some replications were not able to provide estimations for everyone when the IRT 2PL estimation procedure was used. There was only one replication in the gaming dataset that resulted in this, but 68 of the 6000 replications in the educational dataset across all twelve conditions resulted in at least one participant not getting an ability estimate. When this happened, the correlation for that replication was not calculated since some participants' thetas would be missing and those replications were not included in the above table. The IRT 2PL was the only statistical model that results in this and all of these replications were with the 50 sample size. With the 50 sample size, IRT 1PL and 2PL data generation, and the three item conditions, this resulted in 3,000 replications for that sample size meaning only 2% of the replications resulted in the IRT 2PL estimations not being able to calculate an estimate for every person. This probably had minimal impact on results but is important to keep in mind when considering non-Bayesian IRT 2PL models with small sample sizes. Figure 4 focuses on only 2PL data generation and looks at the trend of model estimations by sample size and item size. It can be seen that increasing sample size for the 2PL generated data only resulted in higher correlations with true ability for the IRT 2PL and Bayesian IRT 2PL models.

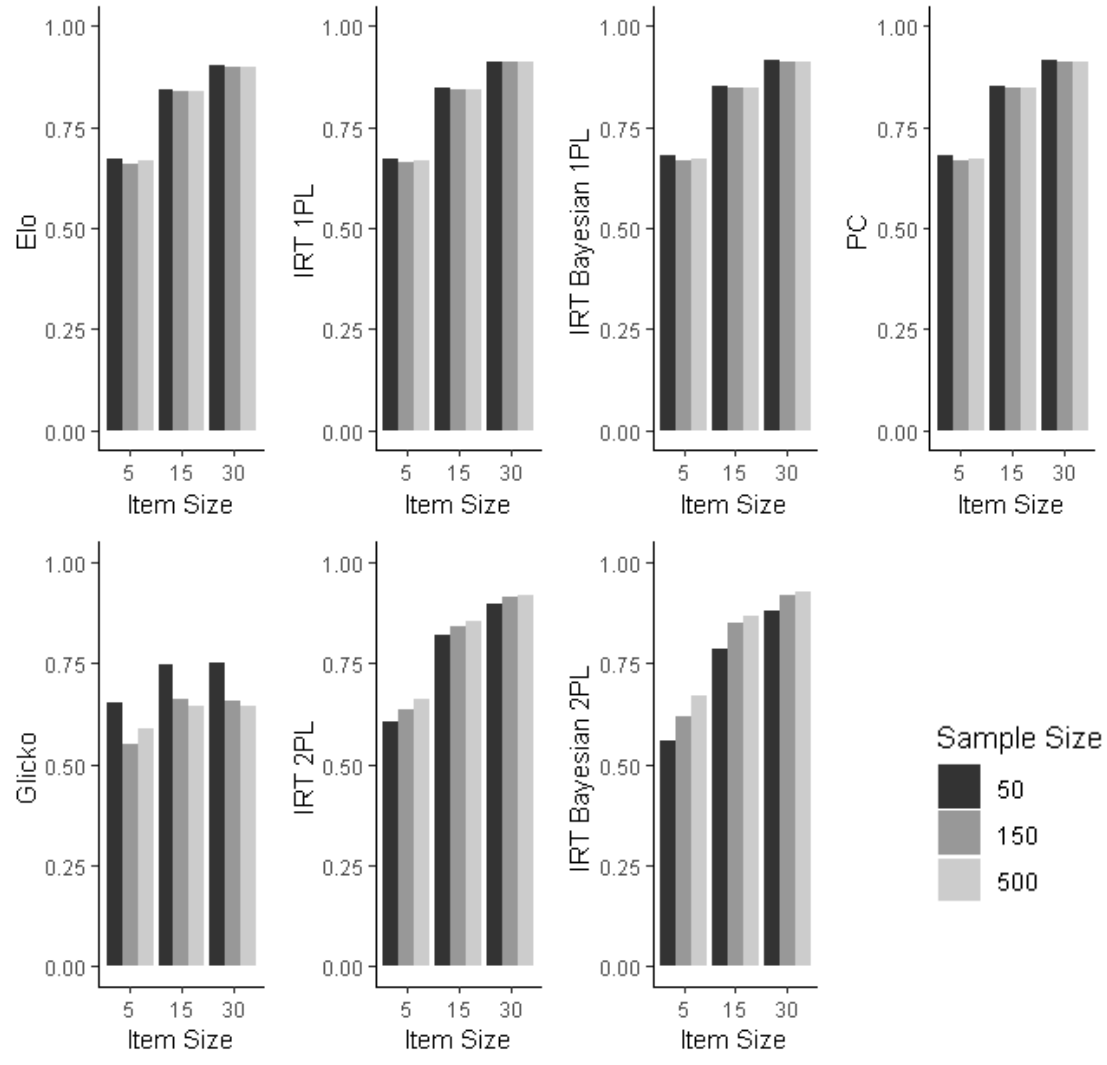

Figure 4. 2PL data generation correlations by item and sample size

The standard deviations of the correlations were also compared and similar to the previous gaming datasets, it appears that sample size shows more of an impact with the standard deviations of the correlations than with the means of the correlations. While for most statistical models sample size seemed to have minimal impact on the correlation between true and estimated ability, nearly all statistical models had lower standard deviations of those correlations with an increase in sample size. This can be observed in Table 8.

Table 8.

Standard Deviations of the Correlations for 1PL and 2PL Generated Data

	Data Generation Method											
	IRT 1PL						IRT 2PL					
	Sample Size						Sample Size					
	50			150			50			150		
	Number of Items			Number of Items			Number of Items			Number of Items		
Model	5	15	30	5	15	30	5	15	30	5	15	30
Elo	.077	.038	.022	.046	.022	.014	.096	.046	.027	.073	.033	.019
Glicko	.081	.062	.059	.051	.045	.046	.105	.083	.079	.101	.091	.086
Prop. Correct	.076	.036*	.019*	.044*	.020*	.012	.094	.043*	.023*	.072*	.031*	.016*
IRT 1PL	.077	.037	.020	.045	.021	.013	.094	.044	.024	.073	.033	.016
IRT 2PL	.119	.069	.038	.068	.024	.014	.141	.065	.040	.102	.037	.017
Bayesian 1PL	.076*	.037	.019	.045	.021	.012*	.094*	.043	.023	.072	.032	.016
Bayesian 2PL	.137	.091	.066	.078	.030	.015	.156	.084	.050	.113	.041	.016

Notes. Lowest standard deviation for each condition is indicated by a *.

Root Mean Square Error (RMSE) for Gaming Data

In addition to correlations, the average root mean square errors were compared across the study conditions. Root mean square error (RMSE) is an aggregated measure of how far away the estimated statistic is to the true statistic with a smaller RMSE indicating greater accuracy (Crocker & Algina, 1986). Since the correlations are correlations between true and predicted ability and the RMSEs are just another measure looking at the difference between estimated and true ability, the RMSEs and correlations were very highly negatively correlated with higher errors meaning lower correlations between true and estimated ability. In this study, the correlation between the true and predicted correlations and the RMSEs were around -.98 to -.99. Table 9 shows the average RMSEs for the Glicko and Elo data.

Table 9

RMSE for Elo and Glicko Generated Data

	Data Generation Method					
	Elo*					
	50			150		
Model	5	15	30	5	15	30
Elo	.836	.587	.463	.848*	.589	.455
Glicko	.846	.646	.608	.869	.718	.695
Prop. Correct	.866	.588	.448	.882	.591	.446
IRT 1PL	.898	.579	.438	.917	.582	.429
IRT 2PL	1.093	.717	.513	1.101	.727	.513
Bayesian 1PL	.834*	.568*	.434*	.851	.567*	.423*
Bayesian 2PL	1.017	.805	.633	1.024	.810	.624

Notes. Lowest RMSEs for each condition is indicated by a *.

The Elo and Glicko data had nearly identical RMSEs which follows the trend with the correlations between true and estimated ability. The patterns of the statistical models were also similar to what was observed for the correlations with the Bayesian 1PL being the most accurate, that is having the lowest errors, in nearly all conditions.

Root Mean Square Error (RMSE) for Educational Data

The RMSEs for the educational generated data are provided in Table 10. The RMSEs for the IRT generated data tend to be smaller than for the gaming generated data, following the inverse relationship as seen with the correlations. The RMSEs for the Glicko estimations and 2PL estimations with small items sizes tended to be the largest, and the RMSEs for the Elo, proportion correct, and 1PL models are similar, overall. The 2PL models especially had smaller RMSEs with the educational data than with the gaming data.

Table 10

RMSE for IRT 1PL and IRT 2PL Generated Data

	Data Generation Method											
	IRT 1PL						IRT 2PL					
	Sample Size						Sample Size					
	50			150			50			150		
	Number of Items			Number of Items			Number of Items			Number of Items		
Model	5	15	30	5	15	30	5	15	30	5	15	30
Elo	.799	.568	.453	.806	.564	.446	.810	.571	.458	.819	.571	.454
Glicko	.826	.705	.698	.879	.811	.807	.836	.712	.703	.940	.816	.823
Prop. Correct	.794	.557	.430	.797	.549	.420	.804*	.560	.435	.810*	.556	.427
IRT 1PL	.798	.562	.432	.802	.555	.420	.810	.566	.436	.816	.564	.428
IRT 2PL	.885	.631	.488	.843	.576	.435	.885	.610	.472	.846	.563	.419
Bayesian 1PL	.793*	.555*	.427*	.797*	.547*	.415*	.804	.559*	.432*	.810	.555	.424
Bayesian 2PL	.979	.772	.602	.914	.587	.437	.933	.657	.500	.865	.553*	.408*

Notes. Lowest RMSEs for each condition is indicated by a *.

The RMSEs seem large especially considering the magnitude of the correlations between true and estimated abilities. However, these errors are similar in magnitude to other research that has looked at IRT with smaller sample sizes and item sizes (Foley, 2010). Figure 5 shows the RMSEs across data generation types by match/item sizes.

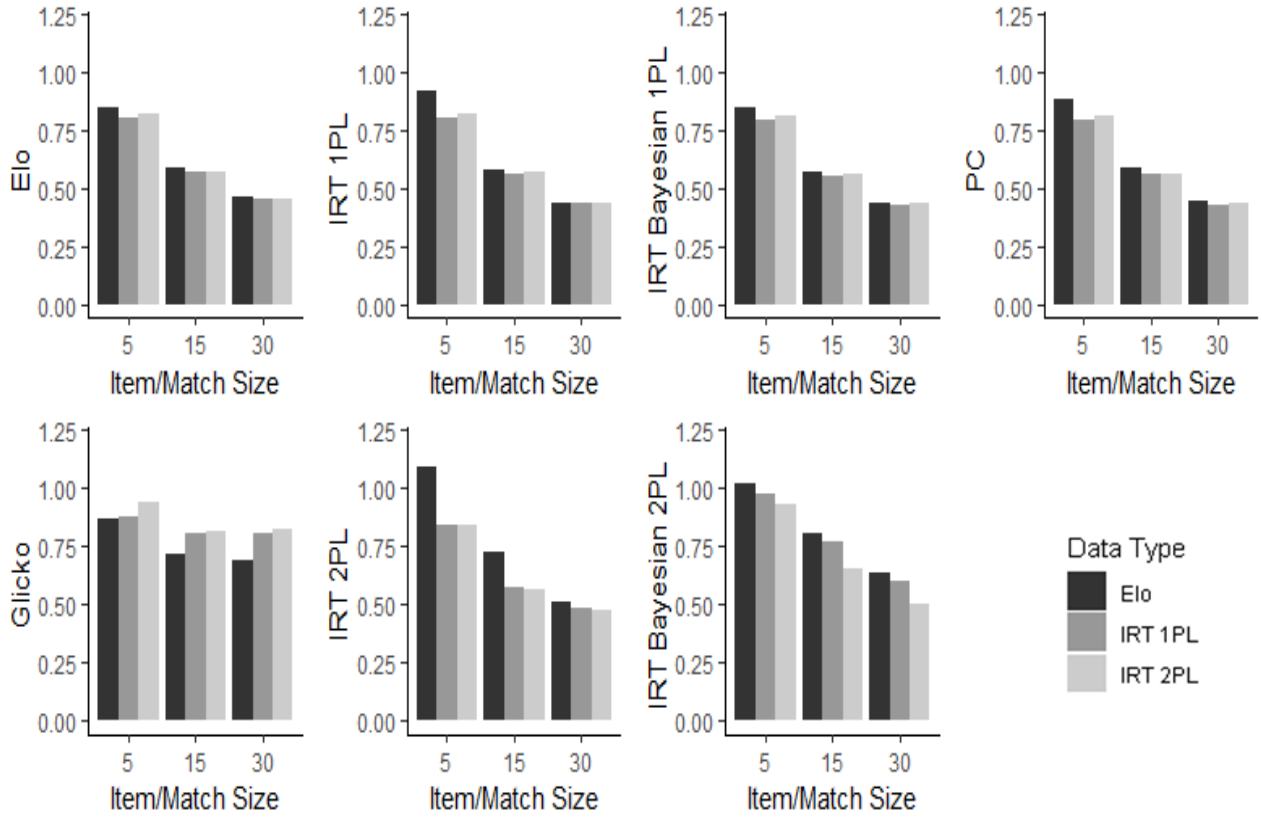

Figure 5. RMSE by data type and match size aggregated across sample size

Computation Time Analysis

The time it took to calculate the estimations for the models were also compared. Though there are many limitations on these numbers, such as different packages, computers, and programs may result in different numbers, they can be a good benchmark to see if the extra time of some of these statistical models may be worth the extra accuracy. The Elo and Glicko data generation were averaged as were the IRT 1PL and 2PL data generation as they had similar times. The average times are provided in Table 11.

Table 11

Average Number of Minutes to Run One Estimation Analysis

	Data Generation Method											
	Elo/Glicko						IRT 1PL/IRT 2PL					
	Sample Size						Sample Size					
	50			150			50			150		
	Number of Matches			Number of Matches			Number of Items			Number of Items		
Model	5	15	30	5	15	30	5	15	30	5	15	30
Elo	.001	.002	.005	.002	.006	.013	.002	.004	.008	.004	.016	.031
Glicko	.001	.002	.006	.003	.008	.016	.002	.005	.010	.005	.020	.037
Prop. Correct	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
IRT 1PL	.007	.005	.006	.012	.008	.009	.009	.004	.004	.007	.004	.005
IRT 2PL	.244	.237	.194	.365	.352	.324	.139	.125	.138	.044	.049	.118
Bayesian 1PL	.216	.582	1.138	1.552	6.039	12.308	.117	.312	1.128	0.303	.909	2.429
Bayesian 2PL	.641	1.638	3.218	3.934	14.275	29.246	.338	.962	3.014	0.937	3.346	8.854

Elo, Glicko, and proportion correct tended to run very fast on all conditions as did the IRT 1PL. While the IRT 2PL ran in under a minute for all conditions, some conditions went up to a run time of 30 seconds which is minimal for just running the data once, but with multiple analyses it might be a concern. The Bayesian models both often took minutes to run but with the educational data it rarely ran for over 5 minutes and the Bayesian IRT 1PL usually only took a minute. The time issue with Bayesian IRT is very noticeable with the gaming data. With a sample size of 150 and 30 matches it takes nearly 30 minutes to run Bayesian IRT 2PL estimations for just one iteration and around 12 minutes for Bayesian IRT 1PL. While other programs may run these analyses faster, just looking at this data using Bayesian analysis on game format data does not seem feasible when one is expecting to run estimations multiple times even if Bayesian IRT 1PL was often the best choice for gaming data.

Comparison of Data Generation Simulation Results

The educational data tended to result in higher correlations for all statistical models overall, but this effect was especially noticeable with smaller numbers of matches/items. The educational data also resulted in lower RMSEs than the gaming data. Across all data types and sample sizes, an increase in matches or items led to a noticeable increase in correlations, and match or item size seemed equally important for both the gaming and educational data.

For the gaming data, an increase in sample size resulted in slightly lower correlations for all statistical models when controlling for match size. However, for the educational data an increase in sample size only seemed to affect the 2PL models with an increase in sample size leading to an increase in correlations. Otherwise, sample size did not seem to noticeably increase or decrease correlations between true and estimated ability.

Overall, it seems that match or item size affected the correlations between true and estimated ability more than sample size. Other studies looking at the correlations between true and estimated abilities for IRT models have also found that sample size had minimal impact on those correlations. However, these studies did find that the item parameters, not the ability estimates, became more accurate with an increase in sample size (Foley, 2010). Since the gaming statistical models did not have a comparison "item" parameter in which to compare to the IRT models, this study focused only on the ability parameters.

Something surprising was how well proportion correct performed in comparison to the other ability estimation procedures in this study. Proportion correct is a very simple statistical model, but it tended to produce true and estimated ability correlations similar to most of the more complex statistical models. The design of the data and the format may explain part of this outcome and will be discussed in the Discussion chapter. The Elo, proportion correct, and the 1PL models all tended to have similar correlations to each other and tended to be the best models except for the 2PL educational data. Only in the case of larger item/sample sizes for the 2PL data generations did the 2PL models result in conditions where their estimations tended to do the best, especially the Bayesian 2PL. Glicko was consistently poorer than most of the other models when applied to both the gaming and educational data with the exception of the 5 match gaming data in which it was similar to most other estimations.

TIMSS Data Analysis Results

The TIMSS data consisted of a sample size of 494 8th grade students from the American sample that fully completed the 2011 binary outcomes mathematics questions for booklet 8. This included the students' responses to 33 items consisting of 11 numerical operations questions, 12 algebra questions, 5 geometry questions, and 5 data and chance questions. The first step was to

estimate the thetas for respondents using all 33 questions. The theta estimates for the participants were correlated to see how the statistical models relate to each other as shown in Table 12.

Table 12

Correlations between Model Estimations of Theta for 33 TIMSS Mathematics Questions

	Estimation Procedure					
Estimates	Elo	Glicko	Prop. Correct	IRT 1PL	IRT 2PL	Bayesian 1PL
Elo	1					
Glicko	.793	1				
Prop. Correct	.965	.758	1			
IRT 1PL	.956	.771	.986	1		
IRT 2PL	.948	.760	.975	.990	1	
Bayesian 1PL	.965	.767	.998	.995	.983	1
Bayesian 2PL	.959	.756	.988	.985	.993	.990

Notes. All correlations were significant at the p <.001 level.

The correlations for the full 33 questions show that many of the estimations were highly correlated from .948 to .998 with Glicko being the only exception and only correlating with the other estimation procedures from the .756 to .793 range. The IRT estimations were highly correlated to their Bayesian IRT counterparts with the range being .993 to .995. Proportion correct was also highly correlated to the IRT models, more so than Elo.

Theta estimates were calculated for differing numbers of items for comparison to the number of items used in the simulation study (5, 15, 30). The correlations between the estimations by number of items are provided in Table 13.

Table 13

Correlations between Model Estimates of Theta by Number of Items in TIMSS Data

	Elo			Glicko			Prop. Correct			IRT 1PL			IRT 2PL			Bayesian 1PL		
	# of Items			# of Items			# of Items			# of Items			# of Items			# of Items		
Estimation	5	15	30	5	15	30	5	15	30	5	15	30	5	15	30	5	15	30
Elo	1																	
Glicko	.871	.819	.802	1														
Prop. Correct	.975	.959	.964	.839	.777	.774	1											
IRT 1PL	.969	.946	.954	.845	.775	.792	.992	987	.986	1								
IRT 2PL	.894	.935	.946	.742	.788	.771	.934	.971	.973	.935	.985	.987	1					
Bayesian 1PL	.975	.958	.964	.840	.779	.786	.998	.999	.998	.994	.993	.995	.927	.977	.981	1		
Bayesian 2PL	.820	.949	.958	.639	.793	.767	.859	.983	.986	.853	.976	.982	.875	.989	.993	.858	.983	.988

Notes. All correlations were significant at the p <.001 level.

A comparison of the correlations across number of items for the empirical data generally resulted in similar correlations for varying item lengths, however there were some exceptions. There were lower correlations for the five item conditions for both 2PL models when compared to other models' ability estimations. Conversely, Glicko had higher correlations for five items, as compared to more items, when compared to other models' estimations except when correlating with 2PL models. This same trend occurred for Elo estimations, however to a smaller degree. Overall, Glicko tended to have the lowest correlations with other statistical models.

It was expected that with an increase in items, all correlations would be correlated more highly with each other since theoretically they should all be more accurately correlated to the unknown true ability. However as mentioned, there were some instances where an increase in items led to some statistical models becoming less correlated with other statistical models. Elo estimations tended to increase in relationship with an increase in items for the IRT 2PL and Bayesian IRT 2PL models, going from .820 to .958, but item size does not seem to affect Elo correlations with IRT 1PL and Bayesian IRT 1PL models with values around .954 to .975. The IRT estimates all seem to increase in correlations with other statistical models as item size increases, going from .935 to .995 for most correlations, except for Glicko, but this increase is especially noticeable between the IRT 2PL and the Bayesian IRT 2PL models where they went from .875 to .993. It is worth noting how highly correlated proportion correct is to the Bayesian IRT 1PL model with correlations ranging between .998 to .9998 and indeed, proportion correct correlated highly with most IRT models.

The second step in the empirical data analysis was to use a set of responses to predict performance on individual items. Two sets of three comparison items were used with one set being three numerical items and one being three algebra items. The questions chosen were based

on the percent of people who got the item correct with one item having around 75% of the participants answering correctly, one with 50%, and one with 25% correct answers. Items of different difficulties (and different content area) were selected as the ability to predict may differ based on the match between the item difficulty and the participant's ability level. Use of multiple items should also help show trends as using a single item may not be a consistent measure of prediction accuracy. Table 14 shows in more detail how the easy, medium, and hard difficulty items compare across the two sets.

Table 14

Percent Correct for Items Used in Prediction Analyses

	Difficulty Level		
Model	Easy (75%)	Medium (50%)	Hard (25%)
Algebra	71.66%	48.99%	25.10%
Number	75.71%	48.18%	29.96%

The estimates associated with the three items with both of the full datasets were used when calculating the probability of a correct answer for that item using the participants' theta estimates after either 5, 15, or 30 items. The 5, 15, and 30 items were chosen while trying to keep the proportion of category items (numerical, algebra, geometry) similar while still being randomly selected. This process was completed with both of the algebra and numerical prediction item sets. For calculating the probability of a correct answer, the IRT 1PL formula was used with all estimates and the estimated thetas were standardized to a z-score distribution.

The kappa coefficient was used to calculate level of agreement between the prediction and the actual outcome of whether the participant got the item correct or incorrect. The kappas

found in this study tend to be weak possibly due to applying the procedure to a comparison of how predicted item and match outcomes correspond to actual match and item outcomes for single items and single matches rather than with composite test scores. Kappa was chosen because it provides more information than using percent agreement, which does not account for the proportion of correct predictions that would be expected based on a random outcome. The kappa coefficients for the algebra questions are shown in Table 15 and the kappa coefficients for the numerical questions are shown in Table 16.

Table 15

Kappa Coefficients for Three Algebra Questions by Number of Matches used for Estimation

	Difficulty Level								
	Easy (71.66%) Number of Items			Medium (48.99%) Number of Items			Hard (25.10%) Number of Items		
Estimations	5	15	30	5	15	30	5	15	30
Elo	.034	.012	.054	.149	.234	.194	.264	.292	.312
Glicko	.064	.140	.166	.155	.225	.095	.247	.277	.314
Prop. Correct	.191	.359	.434	.141	.186	.178	.254	.356	.395
IRT 1PL	.191	.317	.339	.141	.228	.205	.160	.249	.310
IRT 2PL	.041	.022	.034	.207	.305	.272	.241	.257	.233
IRT Bayesian 1PL	.191	.317	.399	.141	.228	.205	.190	.303	.310
IRT Bayesian 2PL	.239	.354	.385	.150	.224	.241	.246	.316	.320

Table 16

Kappa Coefficients for Three Numerical Questions by Number of Matches used for Estimation

	Difficulty Level								
	Easy (75.71%) Number of Items			Medium (48.18%) Number of Items			Hard (29.96%) Number of Items		
Estimations	5	15	30	5	15	30	5	15	30
Elo	.095	.272	.308	.199	.245	.283	.030	.034	.038
Glicko	.122	.322	.218	.152	.249	.260	.153	.213	.255
Prop. Correct	.099	.269	.367	.297	.354	.372	.264	.366	.383
IRT 1PL	.043	.186	.239	.297	.341	.348	.264	.216	.287
IRT 2PL	.057	.237	.264	.252	.343	.355	.211	.234	.324
IRT Bayesian 1PL	.043	.186	.239	.297	.341	.383	.264	.216	.311
IRT Bayesian 2PL	.043	.239	.273	.235	.353	.341	.211	.234	.324

While there are some differences among questions, overall the accuracy of predictions tended to increase as number of items used in the prediction increased for most of the models. There were some statistical models that had a few items where the kappa was consistently very low. These were the Elo and IRT 2PL models with some items only having a kappa between .012 and .054 no matter the number of items used. There does not seem to be consistent differences in kappa coefficients when varying by item difficulty. The statistical model that tended to result in the largest overall kappa coefficients appeared to be the proportion correct model. The next best models based on the kappa coefficients were both of the Bayesian IRT models and then the IRT models though they all tended to have similar sized kappas.

Gaming Data Analysis Results

The gaming data was a dataset with over 20,000 matches and over 14,000 players. It was essentially a snapshot of games from the Lichess online gaming site that was put on Kaggle (J, 2017). The number of matches for the players present in the dataset was positively skewed with most players having only 1 or 2 games in the dataset. The analysis focused on 112 players who had at least 31 matches with non-duplicating players. The 6^{th}, 16^{th}, and 31^{st} matches of these players were used when calculating the kappa coefficient. Since a large number of players had few games in the system, the opponents' skill level used when calculating the matches' outcome was the skill rating given in the database which would be calculated not only on the matches present in the dataset, but matches that occurred before the dataset. For the kappa analysis, the Elo probability of a win formula was used and scores were rescaled as needed. Similar to the empirical educational data analysis, since some models did not estimate all the parameters needed to use their respective probability of a win/correct answer formula so the focus was more on the thetas generated by the statistical models.

It proved difficult to conduct IRT 2PL and IRT Bayesian analyses on the gaming data due to the nature of how the data need to be formatted for these analyses to be conducted. An approximately 12k by 12k matrix was created for the gaming matches which resulted in memory failure for even one of the university's high performance computers. However, Elo, Glicko, proportion correct, and IRT 1PL were able to be used to analyze the data. The correlations between the ability estimations by number of matches are shown in Table 17.

Table 17

Correlations between Ability Estimations by Number of Matches

	Elo			Glicko			Proportion Correct		
	Number of Matches			Number of Matches			Number of Matches		
Estimation	5	15	30	5	15	30	5	15	30
Elo	1								
Glicko	.990	.964	.931	1					
Prop. Correct	.998	.987	.956	.990	.964	.931	1		
IRT 1PL	.287	.591	.789	.300	.559	.787	.287	.602	.829

Notes. All correlations were significant at the $p < .05$ level.

While some of the correlations between the estimates were consistent across match sizes, the correlations to IRT 1PL thetas from the other estimations increased with match size. Proportion correct was the most correlated to the IRT 1PL model followed by the Elo estimation and these correlations increased with match size. The correlations between Glicko and the Elo and proportion correct models decreased with an increase in match size from .990 to .931. Elo and Glicko had very low correlations with the IRT 1PL model at the 5 and 15 match sizes but correlated at .789 and .787 with 30 matches. Elo and proportion correct were highly correlated but the correlation seemed to decrease with an increase in match size.

Since the estimated thetas were used with the Elo model for the gaming data, scores needed to be rescaled to the respective models being used to compare the participants' ability estimates. When rescaling there were two options on how to rescale. One was to rescale the 112 participants' thetas based on the estimates for the whole sample of around 12,000 players or to rescale using just the 112 participants' values. The means and standard deviations for the different estimates for the whole sample and for just the 112 are shown in Table 18.

Table 18

Mean and Standard Deviations for Rescaling Estimations for Full Sample and Restricted Range

	Number of Matches for Estimation			
	Full Sample Rescale		112 Sample Rescale	
Estimations	Mean	Standard Deviation	Mean	Standard Deviation
Elo	.046	29.760	-24.472	130.244
Glicko	5.437	181.847	-38.112	193.419
Prop. Correct	.518	.475	.466	.174
IRT 1PL	-.121	1.380	-.171	.896

Notes. Numbers were from the estimates for 30 games.

Since neither rescaling method seemed ideal, kappa coefficients were calculated using both versions for rescaling the 112 participants' thetas. The trends tended to be fairly consistent with both versions, especially for larger match sizes though some statistical models seemed more affected by the rescaling method than others. The kappas for the gaming data for both rescaling options are shown in Table 19.

Table 19

Kappa Coefficients by Number of Matches

	Number of Matches for Estimation		
	5	15	30
Estimations	Full Sample, *N* = 112	Full Sample, *N* = 112	Full Sample, *N* = 112
Elo	.055 .122	.134 .156	.258 .258
Glicko	.033 .153	.137 .230	.265 .268
Prop. Correct	.023 .103	.236 .175	.324 .323
IRT 1PL	.176 .179	.125 .164	.359 .342

Notes. Numbers on the left are rescaled using the complete gaming dataset of over 12,000, while numbers on the right are the restricted rescoring (rescaling using only the N = 112 participants with at least 31 matches).

For almost every estimation, the kappa coefficient increases along with number of matches with estimations going from as low as .023 in the 5 match condition to as high as .359 for the 30 match condition. The IRT 1PL model tends to outperform the other estimations at every level of match sizes though the middle match size of 15 shows IRT 1PL actually doing worse than it did with the 5 matches. Depending on the rescale method, Glicko and proportion correct did better than IRT at this level as well. The IRT 1PL outperforms the other estimations at 5 matches with the IRT 1PL kappa being .18 with the next highest being .153. Glicko tended to do better here than in other cases, depending on the rescale option, but the reader is reminded that the opponents' theta was taken from the assigned theta in the dataset which is based on Glicko 2 (J, 2017). It is possible that this could result in a positive bias in favor of the Glicko model.

Proportion correct outperformed both the gaming estimations overall. Even with some matching between participants being present in this data set, proportion correct performed well

even though it only outperformed the other estimates when the thetas for the players were rescaled using only the 112 players. The IRT and Elo models seemed less susceptible to changes in their kappas based on the rescaling option which could be considered when selecting the best model.

Comparison of the Real Data Results

There are factors to consider when comparing the educational and gaming real data sets. The format of the data for the real data analysis is different, even more so than the simulation data. The educational data focuses on participants who answered all the questions that were being analyzed and is a complete data matrix. The gaming data focuses on only 112 participants that had at least 31 unique game matches though thousands of their opponents were including in the estimation process. However, all players do not play against all other players, making the data matrix relatively sparse. Further, all players were included in calculating theta estimates meaning that players that may only have had 1 or 2 matches were included. While this is a common occurrence in real gaming data, the fact that real gaming data is much messier than real educational data could led to some issues when comparing the analyses. The biggest limitation is that three of the model estimations, IRT 2PL and the Bayesian IRT 1PL and 2PL models, were not able to be used for conducting the gaming data analysis. Since the gaming data had to be transformed into a matrix of n x n where n is the number of players, this results in very large matrices that available computers were not able to analyze.

Using analyses that were possible, the correlations between the theta estimates' for the statistical models were much higher for the educational data across all item sizes and were frequently in the .8 and .9 correlation range. The correlations between the theta estimates were much weaker for the gaming data, especially at lower match sizes. Additionally, match size

seemed to affect the correlations between the estimates more for the gaming data than for the educational data. Proportion correct tended to correlate to the other estimates a bit differently based on data type. For the gaming data, proportion correct tended to be most correlated with Elo but for the educational data it was the most correlated with the IRT estimates.

With the educational data, Elo, and sometimes Glicko, tended to be as accurate as IRT 1PL in quite a few cases though there were still issues with Elo being less stable with certain items. When using the kappa estimates, IRT 1PL tended to be the best model across almost all match sizes with the gaming data. The kappa coefficients were similar in value by match and item sizes for the two data sets though the educational data tended to have higher kappas when compared to the gaming data especially with the 15 match and item size. While the educational data had higher correlations between the statistical models' thetas than the gaming data set's estimations, the kappas for most of the models tended to be relatively similar and fairly weak. This could be due to the instability of using just one item or match for an outcome when calculating a kappa coefficient.

Summary

The simulation and real data analysis had many similar trends. The generated educational data and the empirical educational data tended to have better correlations and predicted outcomes than the gaming datasets. The increase in match and item sizes led to better correlations and kappas with the simulation and real data analysis with both types of data. The Glicko tended to correlate poorly to the true ability in the simulation and to the other estimations in the real data analysis but the kappa estimates for Glicko did not seem to do as poorly and were often similar to other estimates. The 2PL models did much better with larger sample sizes and item/match sizes though this model provided better estimates with the educational data than with the gaming

data. The 1PL models, Elo, and proportion correct tended to have similar correlations and kappas in both the real data and the simulations for both types of data. Overall, there were differences in the performances of the estimations depending on data type and condition and this knowledge could help determine the usefulness of these estimations across the education and gaming fields.

CHAPTER 5

DISCUSSION

The purpose of this study was to compare how a selection of education and gaming statistical models function for two types of data with small sample sizes and small numbers of items or matches. This study also sought to expand on previous research that has been conducted on comparing educational and gaming statistical models. By having both simulated and real data, comparisons were possible for both known conditions and real-world estimations. In an attempt to further conclusions drawn from the real data analysis, kappa coefficients were calculated to attempt to measure accuracy in prediction in addition to correlations between estimations.

For the discussion, the simulation results will be discussed first, then the empirical data analyses, and then a comparison of the statistical models under the different conditions. Finally, limitations to the study will be discussed along with recommendations for future research.

Simulation Study Aggregated Results

When comparing the aggregated data by data generation methods, Glicko and Elo produced similar correlations between true and estimated abilities. Glicko estimation includes a deviation parameter when estimating ability, and this deviation was included in the Glicko data generation. However, since matrices of matches were being generated based on the true ability, it seemed illogical to assign a large rating deviation for generating the results when we know the true ability. The starting rating deviation was therefore selected to be small, and it stayed small throughout data generation and thus the generated Glicko data ended up being extremely similar to the Elo generated data. A comparison was made of generated Glicko data when the original rating deviation was set to be larger, and the increased error in the generated data resulted in similarly decreased correlations for all statistical models for Glicko data.

The educational generation models consistently had higher correlations between estimated and true ability across all statistical models. This could be due to educational data generation being much cleaner, that is, every participant answered every question leading to a complete matrix of data. In contrast, the gaming data did not have every player face every other player leading to incomplete matrices. Using the average number of matches for the players in the gaming data as compared to educational data where all participants complete the same number of items may have also contributed to the lower correlations in the gaming data.

There were many similarities between both types of gaming and education data when it came to the statistical models. Glicko consistently did poorly for both types. Elo, proportion correct, and both the 1PL models tended to produce similar correlations between estimated and true ability values. While the trends for the statistical models for both types of data were mostly similar, the 2PL model seemed to do poorly on the gaming data, even doing worse than the Glicko model. It could be that the use of the 2PL model with the sparse gaming data matrices is not effective with the small numbers of items and small samples used in the study or that the gaming data does not have a parameter similar to the discrimination parameter the 2PL models try to estimate.

Correlations between true and estimated abilities disaggregated by match/item size were also compared. Correlations consistently increased as match/item size increased with the difference between the 5 and 15 match/item conditions being larger than the 15 and 30 match/item conditions. This diminishing return on number of items added is consistent with previous literature (Crocker & Algina, 1986). However, an increase in sample size did not seem to impact the correlations when controlling for match/item conditions. These correlations and

sample size findings were similar to studies using an IRT 3PL model to estimate ability with small samples and items (Foley, 2010).

Simulation Study with Gaming Data

Data were disaggregated by data generation procedure, sample size, and match/item size to make more specific conclusions. When comparing across gaming data conditions, an increase in matches led to an increase in correlations across all model estimations. While sample size did not increase correlations when aggregated across model generation methods, for the gaming data, some estimations seemed to be negatively impacted with lower correlations when there was a larger sample size. The reason for this may be due to the gaming data format. Since gaming data is paired comparisons among players, more overlap between players can result in better comparisons between them. If you think of gaming data as being a giant web where lines connect players who have been matched together, making comparisons is easier when there are more connections between the players. When sample size increases while holding the number of matches consistent, there is less overlap between the players' matches and it makes it harder to find the connections between them.

Looking at the model estimation results from the simulated data, the 2PL models do not seem to do well with the gaming data. The Glicko model, an estimation procedure for gaming data also did not produce estimates that correlate highly with true ability. There may be many reasons for why the Glicko had poor estimations for the gaming data. The Glicko estimation adds an additional variable where the player's rating changes as a function of both their own and their opponent's "rating deviation". As described, rating deviation is a standard deviation assigned to the player's rating to reflect how sure one is of this rating. This rating deviation goes down as number of matches increases. Since the simulation data tries to have all players have a similar

number of matches, this added variable may not be useful in the design of this study. Another reason may be due to this data not having a growth parameter set. Both Elo and Glicko were designed to accommodate changes in ability levels over the testing period while IRT models were not (Glickman, 1995). It is possible that Glicko may perform better when the data have true ability levels of the participants change over time.

Additionally, another reason the Glicko may be underperforming is due to the parameters chosen to run the Glicko estimation. Identifying the ideal parameters to use for estimating ability can vary based on the type of gaming data being analyzed (Pelánek, 2014). Therefore, the default mean and standard deviation typically used for Glicko and Elo estimation with chess data were used in this study. It is possible that different parameters may have led to an increase in performance of the Glicko estimations but identifying those parameters with certain datasets is a study in itself and was not pursued in this research.

While an increase in sample size did not increase the correlations for the gaming data when item sizes are held constant, the standard deviations of the correlations from the replication got smaller as sample size increased. This means that an increase in sample size did not lead to better correlations, but the correlations of the replications calculated were more stable.

Overall, the best statistical model for the gaming data seems to be the Bayesian IRT 1PL model. At 5 matches, Elo is the best model, though Elo and Bayesian IRT 1PL were very similar at 5 matches. From 15 matches to 30 matches, Bayesian IRT 1PL was the better statistical model though it was very similar to IRT 1PL with 30 matches. These results may demonstrate that the Bayesian IRT 1PL has advantages when being used with smaller match sizes over just the 1PL model, and that the Bayesian IRT 1PL statistical model may improve gaming estimations as compared to Elo. While the proportion correct model was very similar to the Bayesian IRT 1PL

with 5 matches on the gaming data, proportion correct is limited when wanting to match players together. Ideally, when selecting players for matches in online gaming situations, the ideal win rate for all players should be around 50% meaning that people across all abilities win and lose at a relatively equal pace. If proportion correct is very similar for all players, then it would be difficult to see which players really have higher (or lower) abilities. The normally distributed ability simulated in this study is a factor that probably led to proportion correct performing relatively well, but it is important to keep in mind that the other models may be needed when working with less normally distributed data.

Simulation Study with Educational Data

For the educational data generation there were many similar trends. Primarily, an increase in items led to an increase in correlations across all statistical models. The results further indicate that ability estimations with 5 items are relatively poor. To achieve a correlation of .90 or greater, 30 items appear to be needed with the samples used in the study. Similar to the gaming data, though most of the correlations do not seem to increase with an increase in sample size, the increase in sample size led to a decrease in the standard deviation of the correlations meaning that the correlations were more stable. Although the correlations between true and estimated ability did not seem to be related to the sample size in most cases, that does not mean sample size is not important. This study did not investigate the item parameter estimations. However, other studies which focused on IRT analysis with small sample sizes and item sizes looked at not only the correlations between true and estimated ability for IRT analysis, but on the item parameters as well (Foley, 2010). While that study also found that the correlations between true and estimated ability did not show a meaningful increase with an increase in sample size, the estimated item parameters became closer to their true item parameters with an increase in sample

size. The magnitude of this effect was similar to the effect that an increase in items had on the correlations between true and estimated ability (Foley, 2010).

The Bayesian IRT 1PL model tended to be the best statistical model across all numbers of items studied with the 1PL generated data. The Glicko model did poorly with both the IRT 1PL and IRT 2PL generated data. The IRT 2PL models did much better on the educational data than the gaming data. The IRT 2PL and Bayesian IRT 2PL models were the best estimations with the large item sizes (30) and large sample size conditions (150 and 500) with the Bayesian IRT 2PL being the best. The IRT 2PL models were also the only statistical models where an increase in sample size led to an increase in correlations. For instance, for the IRT 2PL generated data, even with the 30 items the IRT 2PL models had the lowest correlations compared to all the other models besides the Glicko with a sample size of 50. However, when sample size increased to 150, the IRT 2PL models became the most accurate. The correlations of the IRT 2PL models increased even more when comparing the 150 sample size to 500 sample size. If the data are assumed to be IRT 2PL data and there is an adequate number of items and a large enough sample size (e.g., 150 or larger), the IRT 2PL models, particularly the Bayesian IRT 2PL, seem to be the best choice.

The more advanced IRT 2PL models did not do as well as the IRT 1PL models with smaller sample and item sizes, even with IRT 2PL generated data, which has been found by other research as well. Sahin and Anil (2016) found that with small item sizes, simpler IRT 1PL models have better model fit than IRT 2PL and 3PL models in estimating item parameters. However with 30 items, the IRT 2PL and 3PL models tended to have better model fit than the IRT 1PL model even with moderate sample size (n = 150). Although their study did not provide ability estimates, the trend observed with model fit estimation was similar to this study's finding

that more advanced IRT 2PL models may not perform as well as IRT 1PL models, even on IRT 2PL data, when there is a small number of items and small sample sizes.

Root Mean Square Errors for the Simulation Study

Since both estimated and true abilities were known for the simulation study, RMSEs were computed. These were highly negatively correlated with the correlations between true and estimated abilities and the trends were similar. When comparing the RMSEs for educational and gaming data, the educational data's RMSEs tended to be smaller than the gaming data. This is likely related to the gaming data being messier and less complete than the educational data. The data format and design may also contribute to why the correlations for the educational data were higher than those for the gaming data as a whole. The Bayesian IRT 1PL model also tended to have the lowest RMSEs across most conditions which mirrored what the correlations between true and estimated ability showed.

Empirical Educational Data

One primary disadvantage to using real data is not knowing the true abilities of the participants which would allow for a comparison between estimated and true abilities. The previous research investigating empirical data and comparing gaming and achievement estimations often looked only at the correlations between those estimations (Antal, 2016; Brinkhuis & Maris, 2009; Wauters et al., 2012). This study included those correlations, but also included a prediction set of items to further the comparisons. For the complete 33 questions on the TIMSS mathematics subtest, the estimated abilities from the seven statistical models were correlated with each other. Glicko did not correlate strongly with any other estimations and this follows the simulation study results. The other estimations all seemed relatively close in level of relationship with proportion correct correlating more with the IRT models than Elo. The

correlations of the estimations by item size were also compared. Number of items had little impact on the correlations between most estimations from the different statistical models. Elo and Glicko however both become more correlated with the 2PL models as the number of items increased. This could be due to the 2PL models not estimating very well with a small number of items more so than the other statistical models.

However, the correlation between estimated thetas provides limited information on what is the best model. By using the kappa coefficient to see how well the model predicts a true outcome on an item or a match, there can be an attempt to see which model makes the best prediction. Proportion correct, 1PL, and 2PL models tended to have the largest kappa coefficients. The Bayesian IRT 2PL model tended to have both the largest and most stable kappas. The IRT 2PL model had one question where the kappas were very low no matter the item size. Elo also had a few items where kappa coefficients were low regardless of the numbers of items. Using just one item to look at the kappas may not lead to very stable analyses, and Elo and the IRT 2PL model seemed more susceptible to that. The 2PL models in general provided an overall better performance for predicting the item-level outcomes, however this may be related to the larger sample size. Since sample size was not varied for either of the empirical datasets, the TIMSS data having 494 participants would give the 2PL models better estimations than what was seen with the 50 and 150 sample sizes for the simulation study.

Empirical Gaming Data

More advanced IRT models were not able to be used for the real gaming data. This is due to how large the matrix was for the empirical gaming data set. While there may be other ways to code the data or code the analysis to make gaming data work more optimally with IRT and Bayesian IRT, this was outside the current studies' purview. One important finding is that when

the number of players is too large (e.g., over 150) in gaming data, IRT may not be practical as running IRT on such a large and sparse matrix may take hours. IRT 1PL was the only IRT model able to run with the empirical gaming data without running out of memory while processing the matrix.

With the real gaming data, the IRT 1PL model tended to outperform the other estimates, according to the kappa coefficients, across most match conditions, even the 5 match condition. However, proportion correct was similar barring the 5 match condition. While it did not prove to be practical to run IRT with a large sample size, the results from the simulation indicate that IRT may be useful for gaming data with small samples (e.g., $n = 150$).

Limitations of the Study

There were a number of limitations to this study. One major limitation being that there were not model fit or item parameter estimates calculated. The gaming statistical models were relatively simple and do not provide item-level or model fit parameters, so there was nothing to compare the IRT parameters to. This also led to the limitation that though there are many ways to measure a model's effectiveness, correlation between true and estimated ability was the main focus for the simulation. IRT is often not recommended for such small sample sizes but given the results from this study it is hard to recommend gaming estimations in some of these instances as a suitable replacement. Just because there are not model fit indices for gaming estimations to see how they work with small sample sizes, that does not mean one should assume that those models would fit better than an IRT analysis. The small correlations between true and estimated abilities would warn against their use in certain conditions.

Generation of the gaming data also involved a lot of variability. As mentioned, gaming datasets are very interrelated as they are players being matched to each other. Therefore, coding

the data so that each player was matched exactly 5, 15, or 30 times seemed extremely difficult. As a result, the data were generated having players with an average of 5, 15, or 30 matches with some players having less and some more. Since the IRT packages chosen for IRT analysis did not allow for items where everyone answered an item incorrectly, this meant that there could not be people who lost all matches and with only 5 matches that is very likely. Two false participants were added for the IRT analysis with one answering every question right and one answering every question wrong to get thetas for the rest of the participants. Those "dummy people" were removed before the correlations were calculated between true and estimated ability. However, even though those two people did not seem to alter the thetas of the other participants, it is still adding in extra data in order to conduct the analysis. Basically, generating the gaming data was less clean than ideal.

Although the main improvement of the Glicko formula over the Elo formula is an addition of a standard error to the ability parameter, when Glicko was generated this error was set to be very small. This was because outcomes generated were based on the true ability parameter and having error for an ability parameter that is known seemed illogical. This led to the Elo and Glicko generated data being very similar so this study was essentially looking at gaming data using only Elo generated data. While there has been research on simulating game outcomes based on tournament style data (Aldous, 2017), there seems to be limited research on generating outcomes for more chaotic match ups that would likely be experienced in online gaming. Other gaming data generation methods should be explored.

Another limitation is the use of the kappa coefficient. While its use was an attempt to measure model effectiveness in predicting a dichotomous outcome, there may have been better

estimation procedures to use. The kappa on just one item (or three for the empirical educational data) can lead to substantial variability in the results.

The IRT 2PL and Bayesian IRT models were also not able to be calculated with the empirical gaming data leaving only the simulation study having a gaming dataset being analyzed by those models. Additionally, since some of the statistical models that calculate a probability of a correct answer/win had parameters that were not estimated by the packages chosen, the thetas calculated were input into the simple Elo or IRT formulas depending on the data. While the kappas shown still seem to follow patterns that are in line with how the statistical models did with the simulation, the use of only the thetas from the statistical models being compared in the kappas is limiting. However, their use adds information above and beyond the correlations among the estimates.

Additionally, in this study players were not matched with the same player multiple times as to mirror a testing situation where a participant would not answer the same item multiple times. However, in online gaming, playing against the same player multiple times may happen. For an IRT analysis to be conducted with this type of data, there would need to be additional modifications to either the way the gaming dataset is formatted or what model assumptions are considered flexible.

The Glicko tended to perform poorly when correlating it with true ability though the kappas did not seem as affected by this. Other research has seemed to conclude that the Glicko should be performing better than the Elo (Glickman & Jones, 1999), but the results from this research study do not show that. It could be that the data in this study is not the type where Glicko can outperform Elo or that other constants for the Glicko formula were needed. As mentioned previously, there may be other conditions where the player deviation rating provides

more useful additional information such as situations where the number of matches completed by players varies significantly. Based on the results from this study, the lower performance of Glicko does not seem to be in line with previous research and a clear reason for that has not been identified.

Recommendations and Future Research

While I feel that future research is needed as comparing gaming and education ability estimations has more work to be completed, there are some recommendations that might be made. This study demonstrates that there may be situations in which IRT and Bayesian IRT can help provide better estimates for gaming data, and situations where the use of Elo for estimation might facilitate educational data. The Elo and IRT 1PL models tended to work well with both types of data with different strengths depending on the condition. Overall, at lower item sizes Elo may do marginally better than IRT 1PL and Bayesian IRT 1PL but their performance was similar in both the gaming and educational data. The Bayesian IRT 1PL model tended to do better with larger item sizes and sample sizes but was still similar to IRT 1PL and Elo.

Proportion correct did quite well across many of the conditions. This may be due to the abilities and item difficulties all being normally distributed in the simulation data and therefore, a proportion correct estimate might be expected to perform fairly well given that random samples of items should be an effective sample of the normally distributed difficulties. Further, there was no matching system in the simulated data while matching might be found in real online data or matching may be used in educational data for CAT tests. If matching was the ultimate goal, either with CAT with educational data or matching similarly skilled players against each other with gaming data, proportion correct would not be useful as the goal would be to match persons with other players of similar ability or items having a difficulty similar to one's ability. When

there are situations where a sample size might not be optimal for IRT analysis and both the item difficulty and skills are assumed to be normally distributed, proportion correct may work well enough for a less technical solution.

While a 150 sample size is not large enough to be recommended for IRT analysis based on prior recommendations (Hambleton et al., 1991), the correlations between true and estimated ability tended to be sufficiently large to use it in online gaming estimation of 30 matches and low-stakes educational testing in the situation of 30 items when group performance is of interest. The 30 item condition usually had correlations between true and estimated abilities over .90. Of concern would be the large standard errors for estimating an individual's ability level. It is possible that with enough items, IRT may still be useful in less formal situations even with more modest sample sizes and when group information is of interest.

Gaming datasets may benefit from IRT models as the Bayesian IRT 1PL model tended to be the best for the simulation data and the IRT 1PL tended to do well with the kappas in the empirical data. However, with the way the data needed to be structured, it makes it impractical to use Bayesian IRT analysis with the gaming data with a large sample size. There may be more effective procedures for using IRT with gaming data. Additionally, Elo analyses were conducted very quickly, but the IRT models and especially the Bayesian IRT models ran slowly with the large, sparse matrices of the gaming data. The amount of time required for the Bayesian IRT 1PL analysis with only 15 items and 150 sample size was around 6 minutes and around 14 minutes for the Bayesian IRT 2PL estimations. These times took even longer with larger sample sizes with the Bayesian IRT 2PL model taking nearly 30 minutes to run with a sample size of 150 and 30 matches. The additional time may not be worth the extra accuracy depending on the circumstances. For instance, if estimations needed to be completed only one time, it may not be

an issue, but for online gaming where estimations are being calculated consistently, the additional time may not allow Bayesian estimations to be viable.

Overall, the research investigating how educational data may work with gaming statistical models is sparse and the research applying IRT estimation to gaming datasets occurs even less. It is hoped that these results can facilitate continued work in these fields. Further studies would need to be conducted on how IRT packages can work better with gaming data before recommendations can be made.

REFERENCES

Aldous, D. J. (2017). Elo ratings and the sports model: A neglected topic in applied probability? *Statistical Science*, *32*(4), 616–629.

Ansley, T. N., and Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*(1), 37–48.

Antal, M. (2016). On the use of ELO rating for adaptive assessment. *Studia Informatica*, *58*(1), 29–41.

Baker, F. B. (2001). *The basics of item response theory*. United States: ERIC Clearinghouse on Assessment and Evaluation. https://doi.org/10.1111/j.1365-2702.2011.03893.x

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*(4), 21–33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x

Bolt, D., and Gierl, M. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, *43*, 313–333.

Brewer, B. (2009). *Introduction to Bayesian statistics*. The University of Auckland. Retrieved from https://www.stat.auckland.ac.nz/~brewer/stats331.pdf

Brinkhuis, M., and Maris, G. (2009). Dynamic parameter estimation in student monitoring systems. *Measurement and Research Department Reports*. Retrieved from http://www.rcec.nl/publicaties/overige publicaties/cito_report.pdf

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Campbell, A. R., and Dickson, C. J. (1996). Predicting student success: A 10-year review using integrative review and meta-analysis. *Journal of Professional Nursing*, *12*(1), 47–59. https://doi.org/10.1016/S8755-7223(96)80074-3

Chang, Y., and Davison, M. (1992). *A comparison of unidimensional and multidimensional IRT approaches to test information in a test battery*. Paper presented at meeting of the American Educational Research Association, San Francisco: CA. Retrieved from https://files.eric.ed.gov/fulltext/ED344940.pdf

Coulom, R. (2007). Computing "Elo ratings" of move patterns in the game of go. *ICGA Journal*, *30*(4), 198–208.

Coulom, R. (2008). Whole-history rating: A Bayesian rating system for players of time-varying strength. In *International Conference on Computers and Games* (pp. 113–124). Berlin: Springer. https://doi.org/10.1007/978-3-540-87608-3_11

Crocker, L., and Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston. Retrieved from https://eric.ed.gov/?id=ED312281

Dangauthier, P., Herbrich, R., Minka, T., and Graepel, T. (2008). TrueSkill through time : Revisiting the history of chess. In *Advances in Neural Information Processing Systems*.

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 239–252.

Elo, A. (1978). *The rating of chessplayers, past and present*. New York: Arco Publishing.

Fatta, G. Di, Haworth, G. M., and Regan, K. W. (2009). Skill rating by bayesian inference. *2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 - Proceedings*, 89–94. https://doi.org/10.1109/CIDM.2009.4938634

FIDE Rating Regulations. (2014). Retrieved September 26, 2017, from http://www.fide.com/fide/handbook.html?id=172&view=article

Flateby, T. L. (1996). *A guide for writing and improving achievement tests*. Tampa, FL: University of South Florida.

Foley, B. P. (2010). *Improving IRT parameter estimates with small sample sizes* (Doctoral Dissertation). Retrieved from DigitalCommons@University of Nebraska - Lincoln.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer Science and Business Media.

Gao, F., and Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education*, *18*(4), 351–380.

Gearhart, C., and Kasturiratna, D. (2018). Implementation of Gibbs sampling within Bayesian inference and its applications in actuarial science. *SIAM Undergraduate Research Online*, 219–231.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). *Bayesian data analysis*. Boca Raton, FL: CRC Press. https://doi.org/10.1017/CBO9781107415324.004

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733-807.

Glickman, M. (1995a). A comprehensive guide to chess ratings. *American Chess Journal*, *3*, 59–102. Retrieved from http://www.glicko.net/research/acjpaper.pdf

Glickman, M. (1995b). The Glicko system. *Boston University*, 1–6. Retrieved from http://www.echecsonline.net/joueurs/doc/The_Glicko_system.pdf

Glickman, M. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, *48*(3), 377–394. https://doi.org/dpi001

Glickman, M. (2012). *Example of the Glicko-2 system*. Retrieved from http://glicko.net/glicko/glicko2.pdf

Glickman, M. E., and Doan, T. (2017). *The US chess rating system*. Crossville. Retrieved from http://www.glicko.net/ratings/rating.system.pdf

Glickman, M. E., and Jones, A. (1999). Rating the chess rating system. *Chance*, *12*(5), 21–28. Retrieved from http://www.glicko.net/research/chance.pdf

Gray-Little, B., Williams, V. S. L., and Hancock, T. D. (1997). An item response theory analysis of the Rosenberg self-esteem scale. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/0146167297235001

Hambleton, R., Swaminathan, H., and Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Herbrich, R., Minka, T., and Graepel, T. (2007). TrueSkillTM: A Bayesian skill rating system. *Advances in Neural Information Processing Systems*, *20*, 569–576. Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2007/01/NIPS2006_0688.pdf

How does GameKnot's rating system work? (2017). Retrieved September 26, 2017, from http://gameknot.com/help-answer.pl?question=29

Hvattum, L. M., and Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470. https://doi.org/10.1016/j.ijforecast.2009.10.002

J, M. (2017). Chess game dataset. Retrieved from https://www.kaggle.com/datasnaek/chess

Jones, L. V., and Thissen, D. (2006). A history and overview of psychometrics. In *Handbook of Statistics* (Vol. 26, pp. 1–27). Amsterdam: Elsevier. https://doi.org/10.1016/S0169-7161(06)26001-2

Lasek, J., Szlávik, Z., and Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, *1*(1), 27–46. https://doi.org/10.1504/IJAPR.2013.052339

Lee, S. Y. (2007). Bayesian estimation of structural equation models. In *Structural equation modeling: A Bayesian approach* (pp. 67–109). Hoboken, NY: John Wiley and Sons.

Levy, R., and Mislevy, R. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: CRC Press.

Linzer, D. A. (2013). Dynamic bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, *108*(501), 124–134. https://doi.org/10.1080/01621459.2012.737735

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford: Addison-Wesley.

Lynch, S. (2007). *Introduction to applied bayesian statistics and estimation for social scientists*. Springer Science and Business Media. https://doi.org/10.1198/jasa.2008.s250

McGrayne, S. B. (2012). *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy.* New Haven, CT: Yale University Press.

Micklich, D. L. (2009). Do first mover advantages exist in competitive board games. *Developments in Business Simulation and Experiential Learning*, *36*, 270–276.

Miller, M. D., and Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, *16*(4), 381–388. https://doi.org/10.1177/014662169201600410

Newton-Fisher, N. E. (2017). Modeling social dominance: Elo-ratings, prior history, and the intensity of aggression. *International Journal of Primatology*, *38*(3), 427–447. https://doi.org/10.1007/s10764-017-9952-2

Patz, R. J., and Junker, B. W. (1999). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178. https://doi.org/10.3102/10769986024002146

Pelánek, R. (2014). Application of time decay functions and the Elo system in student modeling. *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, 21–27. Retrieved from https://pdfs.semanticscholar.org/6fa2/b7b57f10e5bdc1a9f03ad482a3c92910a30a.pdf

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers and Education*, *98*, 169–179. https://doi.org/10.1016/j.compedu.2016.03.017

Pickett, K. E. (2001). Multilevel analyses of neighbourhood socioeconomic context and health outcomes: A critical review. *Journal of Epidemiology and Community Health*, *55*(2), 111–122. https://doi.org/10.1136/jech.55.2.111

Plummer, M., Stukalov, A., and Denwood, M. (2018). rjags.

Reckase, M. D., and McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*(4), 361–373. https://doi.org/10.1177/014662169101500407

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research. Illinois. Retrieved from https://cran.r-project.org/package=psych

Robitzsch, A., Keifer, T., and Wu, M. (2018). TAM: Test analysis modules. Retrieved from https://cran.r-project.org/package=TAM

Şahin, A., and Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory and Practice*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Sinharay, S., Johnson, M. S., and Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298–321.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Hoboken, NJ: John Wiley and Sons. https://doi.org/10.1002/0470092602

Steinberg, L., and Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, *1*(1), 81–97. https://doi.org/10.1037/1082-989X.1.1.81

Stephenson, A., and Sonas, J. (2016). R package "PlayerRatings." Retrieved from https://cran.r-project.org/web/packages/PlayerRatings/index.html

Stocking, M., and Lord, F. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

Tett, R. P., Jackson, D. N., and Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, *44*(4), 703–742. https://doi.org/10.1111/j.1744-6570.1991.tb00696.x

The ACT technical manual. (2017). Iowa City, IA: ACT.

TIMSS 2011 publications. (2011). Retrieved from https://timssandpirls.bc.edu/timss2011/international-database.html

Tsang, C. S. C., Ngan, H. Y. T., and Pang, G. K. H. (2016). Fabric inspection based on the Elo rating method. *Pattern Recognition*, *51*, 378–394.

Upcoming tournaments. (n.d.). Retrieved from http://www.uschess.org/component/option,com_wrapper/Itemid,199/

van der Linden, W. J., and Hambleton, R. (1997). *Handbook of modern item response theory*. New York, NY: Springer Science and Business Media.

Veldkamp, B. P., and Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: Avaliação e Pol\'Iticas Públicas Em Educação*, *21*(78), 57–82. https://doi.org/10.1590/S0104-40362013005000001

Véron, M., Marin, O., Monnet, S., Universités, S., Regal, É., and Regal, É. (2014). Matchmaking in multi-player on-line games : Studying user traces to improve the user experience. In *Network and Operating System Support on Digital Audio and Video Workshop* (pp. 7–12). https://doi.org/10.1145/2578260.2578265

Viera, A. J., and Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wauters, K., Desmet, P., and Noortgate, W. (2011). Monitoring learners' proficiency: Weight adaptation in the Elo rating system. *Proceedings of the 4th International Conference on Educational Data Mining*, 247–251. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84857467666&partnerID=40&md5=66495f03fb3937ec71b243b8f4fbca74

Wauters, K., Desmet, P., and Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers and Education*, *58*(4), 1183–1193. https://doi.org/10.1016/j.compedu.2011.11.020

Weng, R. C., and Lin, C.-J. (2011). A Bayesian approximation method for online ranking. *Journal of Machine Learning Research*, *12*, 267–300.