6-15-2024

Capturing Latent Abilities and Latent Capacities of Professional Golfers Using Nonlinear Mixed Effects Growth Modeling

Mac Wetherbee

Capturing Latent Abilities and Latent Capacities of Professional Golfers Using Nonlinear Mixed Effects Growth Modeling

Abstract

This study demonstrates an effective and innovative approach to measuring the latent athletic abilities and capacities of professional golfers. I used nonlinear mixed effects growth modeling (e.g., Dynamic Measurement Modeling) to measure professional golfers' ability levels and capacities for improvement. I accomplished this using a two-stage modeling approach. First, a crossed linear mixed effects model estimated each player's ability level in each year. In the second stage, I used the results from the first stage to estimate several candidate nonlinear growth trajectories for players' abilities over time. The quadratic growth trajectory was the best-fitting of these trajectories and was used to estimate each player's individual-level capacity (maximum predicted ability). Validation results indicate that the ability estimates from stage one can outperform existing unidimensional measures of golfing ability and that the capacity estimates from the second stage are reliable and provide better forecasts of players' future abilities than do single-timepoint estimates. This study demonstrates the applicability of latent variable statistics and longitudinal growth models to the study of sports, provides a novel statistical method to estimate player abilities and capacities in professional golf, and provides a tutorial for estimating Dynamic Measurement Models (DMM) using *R*.

Document Type

Dissertation

Degree Name

Ph.D.

First Advisor

Yixiao Dong

Second Advisor

Nick Cutforth

Third Advisor

Mei Yin

Keywords

Dynamic measurement modeling (DMM), Golf science, Golf statistics, Growth modeling, Mixed effects modeling

Subject Categories

Applied Mathematics | Applied Statistics | Educational Assessment, Evaluation, and Research | Other Applied Mathematics | Physical Sciences and Mathematics | Sports Studies | Statistical Models | Statistics and Probability

Capturing Latent Abilities and Latent Capacities of Professional Golfers using Nonlinear Mixed Effects Growth Modeling

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Mac Wetherbee

June 2024

Advisor: Yixiao Dong

Author: Mac Wetherbee
Title: Capturing Latent Abilities and Latent Capacities of Professional Golfers using Nonlinear Mixed Effects Growth Modeling
Advisor: Yixiao Dong
Degree Date: June 2024

Abstract

This study demonstrates an effective and innovative approach to measuring the latent athletic abilities and capacities of professional golfers. I used nonlinear mixed effects growth modeling (e.g., Dynamic Measurement Modeling) to measure professional golfers' ability levels and capacities for improvement. I accomplished this using a two-stage modeling approach. First, a crossed linear mixed effects model estimated each player's ability level in each year. In the second stage, I used the results from the first stage to estimate several candidate nonlinear growth trajectories for players' abilities over time. The quadratic growth trajectory was the best-fitting of these trajectories and was used to estimate each player's individual-level capacity (maximum predicted ability). Validation results indicate that the ability estimates from stage one can outperform existing unidimensional measures of golfing ability and that the capacity estimates from the second stage are reliable and provide better forecasts of players' future abilities than do single-timepoint estimates. This study demonstrates the applicability of latent variable statistics and longitudinal growth models to the study of sports, provides a novel statistical method to estimate player abilities and capacities in professional golf, and provides a tutorial for estimating Dynamic Measurement Models (DMM) using *R*.

Table of Contents

List of Tables

Chapter One: Introduction

Background

Golf is a sport played, in simplest terms, by trying to hit a ball on the ground with a stick (or 'club') into a small hole in the ground, which is typically hundreds of yards away. The golfer tries to minimize the number of times that he or she hits the ball (called 'shots' or 'strokes'). This process is then repeated 18 times to complete one round of golf.[1] Golfers with greater ability generally require fewer shots to complete the 18-hole round of golf than do golfers with less ability.

For decades, scholars (and golfers and golf coaches) have attempted to measure golfing ability quantitatively. At the amateur ranks (i.e. for golfers not on a professional tour), this is typically done either through the player's average score above or below par or through a handicap rating system. For professional golfers, the average score above or below par is also frequently used (e.g. J. Baker S. Horton & Deakin, 2006), as are more skill-specific measurements, such as average driving distance, percentage of putts made within 20 feet, etc. (e.g. Nix & Koslow, 1991). More quantitatively sophisticated ways of measuring golf performance have also arisen in recent years, improving upon the more traditional methods of measuring golfing ability (e.g. Stigler & Stigler, 2018;

[1] A complete description of how golf is played, including a list and definition of golf-related terms, can be found in Appendix A.

Broadie & Rendleman Jr, 2013; Connolly & Rendleman Jr, 2012; Elmore & Urbaczewski, 2018). However, both the more traditional measurements and the newer, more sophisticated methods suffer from one or more of several potential deficiencies.

Perhaps most notably (from a statistical perspective), most measurements of golf performance and ability are treated as observed variables. Inherently, though, golfing ability is a latent variable. It exists—some golfers have more ability than others—but is not something that humans can easily observe or measure precisely by watching a golfer or by looking at his or her scores. By using observed variables as proxies for the latent ability, researchers and practitioners are making untested assumptions about how effectively those proxies measure the true latent ability. Thus, if golfing ability is a latent variable, it would likely be more effective to treat it as such statistically by using a latent variable statistical model to estimate golfing ability.

Second, many ways of measuring golfing ability force researchers to make the assumption that all golf courses are created equal. For example, if we assume that average driving distance measures golfing ability (or at least a specific skill or dimension of golfing ability such as "power" or "strength"), we assume that those skills would manifest equally across courses. However, not all individuals play the same courses, and different courses are likely to have different characteristics that affect average driving distance. One course might be drier than another, allowing the ball to roll further. A course at a higher altitude would allow the ball to travel further through the air. One course might be narrower than another, encouraging players to sacrifice some distance in favor of accuracy. This might not be overly problematic if all golfers played the same

courses. However, that is certainly not the case, even among professional golfers. No player chooses to play every week—each player chooses which tournaments to play and which weeks to take off. Furthermore, there are many different professional golf tours, and two players on different tours are not likely to play in the same tournaments as each other very often, if at all. Amateur and recreational golfers are even less likely to play the same courses as each other—they are far more likely to play courses near their place of residence than those further away. Thus, by condensing the measurement of the ability down to an average over a given time period, there is potential measurement error and bias.

Third, most existing ways of measuring golfing ability assume that each shot of a given length is equally difficult. For example, if we use the percentage of putts made within 20 feet to represent a golfer's ability (or at least a specific skill or dimension of golfing ability such as "putting ability"), we assume that each player has an equal distribution of difficulty within 20 feet. However, some golfers might have more putts within the lower end of the range (e.g. more putts from one to eight feet) while another might have more putts within the higher end of the range (e.g. 12 to 20 feet). Additionally, even at a given length, some shots are going to be harder than others. A straight uphill putt from 10 feet is likely to be much easier than a downhill 10-foot putt that curves (or "breaks") a foot to the left or right. If these are non-randomly distributed across players, then measuring golfing ability by looking at the observed variable of putting percentage will be misleading.

Fourth, from a more practical perspective, many (though not all) of the measures that are used to measure golfing ability require detailed shot-by-shot data. For example, Strokes Gained, which is the current state-of-the-art in measuring golfing ability, requires one to know how far away from the hole the ball was before the shot, the surface it was on before the shot (fairway, rough, pine straw, sand, etc.), how far away from the hole it ended up after the shot, and the surface on which it ended up after the shot. This is far beyond the level of detail typically tracked by recreational golfers, and it is even beyond the scope of data provided by most professional golf tours. Thus, even if this were the optimal way to measure golfing ability, it may not be feasible in many cases.

Fifth, many advanced statistics across sports require comparison to a reference group such as the "average" player or a "replacement level" player. Strokes Gained follows this same logic: it is measuring the strokes gained above what the average player would be expected to achieve. This requires defining who the average golfer is for each scenario. For a professional golf tour, this may seem straightforward: the average golfer on that tour. Even in that scenario, however, a different average could be chosen: the average player across all professional tours, for example. For recreational golfers, it is even less clear who the reference golfer should be. Is it the average recreational golfer worldwide? The average recreational golfer in his or her country? The average recreational golfer in his or her state? The average recreational golfer with the same handicap? The average recreational golfer of the same age? There is no clear correct answer to this question.

Sixth, some of the ways that golfing ability has been measured impose an implied factor structure on the data. They assume that putting is one ability, chipping and pitching is another, approaching the green is another, etc. These assumptions are based on logical and reasonable theories about the structure of golfing ability, but they are just assumptions. They have not been tested in a systematic way. Because of this, there may be additional dimensions that have been under-studied and under-represented in the study of golfing ability. Similarly, some golfing outcomes that have been presumed to measure different abilities may actually be indicators of the same underlying latent ability.

Finally, most ways of measuring golfing abilities that recognize multiple dimensions treat these dimensions as if they are totally distinct from each other. However, in the way that they are measured as observed variables, these variables are likely to be correlated with each other, and some outcomes may even be measuring multiple abilities simultaneously. For example, one frequently used way of measuring a player's ability from sand traps is his or her sand save percentage (the percentage of the times in a greenside sand trap that a player saves par). Undoubtedly, part of what determines this percentage is indeed how well the player hits shots from the sand trap. However, a player's putting ability likely also affects this statistic, as a player that makes more putts will be able to save par from further distances (or more frequently from the same distance). Thus, untested assumptions about which indicators measure which abilities and how different abilities are correlated with each other may lead to measurement error.

These issues and resulting biased ability estimates can lead to misunderstandings among players, coaches, media members, sponsors, and fans. Players and coaches may identify the wrong strengths and weaknesses and may therefore use their time and training efforts inefficiently. Sponsors may choose the wrong players to sponsor, leading to suboptimal marketing spending. Fans may have unrealistic expectations for their favorite players, and they may have even selected their favorite players using inaccurate information. Members of the media may be inadvertently presenting incomplete and misleading stories, further exacerbating these other problems.

Moving Toward Latent Variable Measurement

In this project, I have addressed many of these issues and have improved on the existing methods of measuring golfing ability. I developed new measures of golfing ability using latent variable statistics applied to the publicly available results of elite professional golfer tours. In addition to measuring a golfer's ability at a fixed time point, I also created a new quantity of interest—professional golfers' capacities—that measures players' capacities for ability in the future. In total, I addressed four related research objectives.

Research Objective 1: To review and synthesize the extant research on estimating golfing ability.

The vast majority of research on the measurement of golfing ability has treated golfing ability as an observed variable (or as a set of observed variables) rather than as a latent ability. However, there have been some recent movements toward using advanced statistical measurement techniques to measure golfing ability. Chapter Two reviews and

synthesizes the extant research on the estimation of golfing ability, including both the observed variable approaches and the latent variable approaches.

Research Objective 2: To estimate the ability levels of professional golfers as a latent ability.

In this stage, I have measured golfing ability among professional golfers as a unidimensional latent ability. I used a crossed linear mixed effects model with random effects for course, date/round, course-year, player, and player-year. The outcome variable was the player's score in the given round. The sum of the random effects for the player and for the player-year thus provide an estimate for how many golf strokes per round above or below average that player is for the particular year. Though we do not always think of mixed effects models as being latent variable measurement models, the conceptualization of a random effect as measuring the unobserved effect of some grouping variable is effectively the same as the measurement of latent abilities in more conventional measurement models.

One additional benefit at this stage is the random effect for course. This random effect provides a measure of how difficult each course is. Although it is not a specific objective of this dissertation, this could provide a more accurate and more cost-effective way to measure course difficulty. Current methods require expert evaluators to visit each course.

This model to measure a professional golfer's ability is similar to the SBSE ("Score-based Skill Estimate") model (Broadie & Rendleman Jr, 2013). The new version in this project included a few additional components: the inclusion of positional

variables, the nesting of a player's ability in a given year within a broader player-ability random effect, and nesting a course-round difficulty measure within a broader course difficulty random effect. Nonetheless, the logic is the same: a player's score in a given round of golf is determined by situational factors (course difficulty, weather, etc.), the player's ability, and random variation.

I validated the ability level estimates produced from the linear mixed effects model and confirmed their accuracy and utility. To do this, I assessed the strength of the correlations between the newly produced ability level estimates, one alternative measure (a pooled, age-invariant measure of ability), and three existing unidimensional measures of golfing ability—the Official World Golf Rankings (OWGR), Total Strokes Gained (the current state-of-the-art unidimensional measure of golfing ability), and the SBSE model. Ideally, the new estimates should be positively but not perfectly correlated with all of these other measures (hypothesized correlations of between 0.6 and 0.8). I also compared the predictive ability (lowest mean square error) of the new estimates to the predictive abilities of the other unidimensional measures of golfing ability. To demonstrate validity, the new ability estimates needed to perform at least as well as the existing measures of ability, and they did so.

Research Objective 3: To assess the shape of longitudinal ability growth of professional golfers.

To accomplish this objective, I utilized the scores for golfers over time produced in Objective 2 above. I then used these longitudinal scores to estimate multiple shapes of nonlinear mixed effect growth curves: two S-shaped growth trajectories, two J-shaped

growth trajectories, and a quadratic growth trajectory. Each of these models was estimated first as a marginal model (i.e., fixed effects only). I then compared the fits of these models using Bayesian Information Criterion (BIC) and mean square error (MSE) to decide which one provided the best explanation for how golfing ability grows over the course of professional golfers' careers. The best-fitting of these trajectories (the quadratic model) was then used in future stages.

Research Objective 4: To estimate the capacity scores for professional golfers.

Dynamic Measurement Modeling (DMM) provides a conceptualization of nonlinear mixed effects models as measurement models, providing estimates of a person's "capacity" for future growth (Dumas et al., 2020;McNeish & Dumas, 2017). In other words, these models take a person's existing scores over time to estimate the maximum possible ability level that the person could achieve in the future in the specific domain. To date, these DMM capacity estimates have primarily been estimated using a random effect on the upper asymptote parameter in S-shaped and J-shaped growth curves, but a similar logic could apply to quadratic growth and other growth trajectories as well. In some of these growth trajectories (such as a quadratic model), the capacity would be represented by the individual-level estimate for the maximum value of the function, while the time at which the function reaches its maximum may also be of substantive interest. Using the quadratic growth trajectory, which fit best in the previous stage, I added random effects to some of the parameters. Most importantly, the player-level random effect for the maximum was used to estimate the capacity value for each golfer.

To validate these results, I ran the DMM model on restricted datasets formed of only early-career ability estimates. In many practical uses, DMM models will be used to estimate capacity scores for individuals before the individuals have reached their capacities. Thus, to be viewed as a valid measurement model, a DMM model should be able to produce useful capacity estimates from only those pre-maximum datapoints. I ran the restricted model using two different age thresholds (both of which are below the model-average age at which players reach their capacities): a dataset of only ages under 25 and a dataset of only ages under 30. I then compared these capacity score estimates to the capacity score estimates estimated by the full model, assessing the extent to which the estimated restricted-model capacity scores deviate from the full-model capacity scores. There should be a relatively strong positive correlation between the restricted-model capacity scores and the full-model capacity scores, and results confirm this strong correlation. Along similar lines, I also used a variation of 10-fold cross-validation to validate the results by assessing the extent to which the results are consistent across subsamples of the data.

The second step in the DMM validation compared the model fit of the final DMM model selected to that of a baseline, naïve linear growth trajectory. Linear growth is not desirable from a DMM perspective because each individual's capacity score would be infinite. However, it is still theoretically possible that linear growth could be the best fitting growth trajectory. To provide evidence that the DMM growth trajectory is valid, I compared model fit statistics (BIC and MSE) for the selected DMM model and

theoretical linear growth. As expected, the DMM model provided a better fit (lower BIC and MSE).

As a final validation step for the DMM model, I compared the capability for DMM capacity estimates to predict future performance to that of other estimates of ability that are captured by single-timepoint estimates. One benefit of DMM capacity estimates is that they reduce the reliance on single-timepoint scores as implicit indicators of future success. This benefit has been shown in educational settings, and I tested whether it holds true in the measurement of golfing ability. In the case of golfing ability, single-timepoint estimates took the form of latent ability estimates at a single age (age 28, or age 25, for example). To demonstrate validity and utility, DMM capacity estimates should be more effective at predicting future ability than any single-timepoint estimates, and results confirmed this.

Contributions

This dissertation makes several contributions. First, it further demonstrates the utility of applying latent variable statistical models to the study of sports. There have been some tepid attempts to move in this direction across sports analytics, but most statistical analyses in sports still rely on observed variables. This dissertation provides a template for utilizing latent variable models and applying them to the domain of sports analysis.

Second, the mixed effects model(s) demonstrate an improved method for estimating golfing ability without using shot-level data. This may be particularly relevant for estimating ability levels for college golfers, recreational golfers, and golfers on

smaller professional tours. Instead of being forced to rely on statistically dubious measures of golfing ability, this new measure provides a statistically valid way to measure golfing ability without requiring onerous data collection requirements. Results indicate that the models in this study also outperform existing unidimensional measures of golfing ability, so this benefit should be relevant even to golfers with greater access to data.

Third, the nonlinear growth model provides the first quantitative method in the literature to forecast future golf ability. This may be of particular interest to sponsors when deciding which young professional golfers to sponsor, to college coaches when deciding which high school players to recruit, and to tournament organizers when deciding which young players to provide with entries into tournaments.

Fourth, the use of DMM capacity estimates provides one of the few examples of DMM being applied to a real-world dataset. Most DMM research to date has been conducted to show the method's efficacy and benefits as opposed to substantive implementations in which the researcher or practitioner is actually interested in the estimates themselves. Thus, this is one of the first studies to use DMM as an extant method of generating meaningful capacity estimates on a dataset of substantive interest—one in which the subjects are not anonymous. This may be a relevant contribution to the DMM literature, helping the method move more into the mainstream of latent variable statistical methods.

Fifth, this is likely the first study to apply DMM outside of the sphere of educational measurement. Because DMM has been developed in the context of

educational measurement, its applications have exclusively been in this realm: reading ability (e.g., Dumas & McNeish, 2018), mathematics ability (e.g., Dumas, McNeish, Sarama, et al., 2019), medical licensing exams (e.g., Dumas, McNeish, Schreiber-Gregory, et al., 2019), etc. By applying this modeling paradigm to a substantive area outside of education, this dissertation demonstrates the conceptual and statistical efficacy of this type of model to other fields of research. In sports statistics specifically, forecasting a player's future growth is of natural interest to many teams, coaches, players, and fans. This study provides a demonstration of a new and improved way to do this in the future.

Sixth, all previously published DMM models have been conducted using *SAS*. I demonstrate that these models can also be conducted using existing packages in *R*. Since *R* is free and open source, the ability to implement these models in *R* should increase access to DMM models, particularly for researchers and practitioners who do not have institutional access to *SAS*. I demonstrated the code used, providing a tutorial on how to run these models in *R* so that readers can use this code as a guide for their own future research.

Seventh, this study continues the expansion of curve shapes that can be incorporated into the framework of DMM. Most previously published DMM studies have primarily focused on either J-shaped or S-shaped growth curves. Both of these growth curve shapes have an upper asymptote parameter that can, once a random effect is added on this parameter, effectively measure an individual's capacity. However, these growth curve shapes inherently assume that individuals never lose the ability being studied. This

may make sense in an education context: individuals in this context are all likely increasing their ability, so it may not make sense to model any decrease in ability. In practice, however, most latent abilities do actually decrease eventually. Individuals' mathematics ability may decrease after high school or college as many adults end up using only certain math skills in their careers. Reading ability may decrease later in life due to cognitive decline. I demonstrated this in the context of golfing ability: we can still estimate an individual's future capacity while also using a growth shape that allows us to model the eventual decline of an individual's skill after reaching his or her peak. I included quadratic growth to demonstrate this possibility.

Eighth, a by-product of the mixed effects model to estimate individual golfers' ability levels is an estimate of each course's difficulty. In fact, this is analogous to the estimate of an item's difficulty in item-response theory (IRT). Just as an IRT model simultaneously estimates each individual's ability and each item's difficulty, the model in stage one of this project estimates each individual's ability and each course's difficulty. This demonstrates a statistically valid way of measuring course difficulty, and it may represent a more accurate and cost-effective way of evaluating courses than the current way that courses are evaluated under the handicap rating system. In this new way of measuring course difficulty, the difficulty value would be determined statistically using the actual observed scores of individuals playing that course.

Ninth, this dissertation provides an alternative to the current handicap rating system for evaluating the abilities of recreational golfers. The current handicap system rates each course based on factors such as length, altitude, number of sand traps, etc.

Each player's handicap rating is then based on how they perform relative to the expectations of those difficulty ratings. However, the mixed effects model in stage one of this study estimates players' abilities and course difficulties simultaneously. Thus, this provides an alternative to the current system: each player's estimated ability level and each course's estimated difficulty rating would be continually updating with every round of golf played. No more trips or subjective ratings by course raters would be necessary.

Finally, the models used in this dissertation allow for the statistical comparison of players across tours and levels. There are many professional golfing tours, and many of the players on these tours never play against the players on other tours. It is therefore difficult to compare the ability levels of these golfers. By linking the tours statistically using any cross-pollination between the tours, both the ability and capacity estimates are on the same scales across tours. This allows for the current ability levels and the future capacity estimates to be compared for two players who have never been on the same continents as each other.

Chapter Two: Literature Review

In the United States alone, the value of golf club memberships sold in the 1990's was over $3,200,000,000. Worldwide, there are tens of thousands of golf courses, and 55 million people play the sport. Academically, there are 11 distinct golf science disciplines and hundreds (perhaps thousands) of academic papers published on the study of golf (Farrally, Cochran & Thomas, 2003). Golf matters to people in the world, and the study of golf matters to scholars.

This literature review proceeds in two broad sections, each with two major subsections. The first broad section is the measurement of golfing ability. Within it, previous research is separated into those studies that have treated golfing ability as an observed ability (or set of abilities) and those studies that have treated golfing ability as a latent ability (or set of abilities) to be measured statistically. The second broad section addresses the statistical methods that are utilized in this dissertation. Within this section, previous research is separated into that on the use of mixed effects modeling broadly and that on the use of Dynamic Measurement Modeling (DMM) specifically.

Measuring Golfing Ability

Golfing Ability as Observed Variable(s)

Using Counting Variables to Measure Golfing Ability

In the context of golf research, Leahy (2014) differentiates between performance assessment and performance analysis. Under this bifurcation, performance assessment research focuses on creating new statistics, measures, or quantities and then re-ranking golfers based on the scores of the newly created values. Performance analysis, on the other hand, seeks to identify which skills or sets of skills are most important in predicting or explaining golfers' performance and success (Leahy, 2014, p. 20-30). They differ in that the performance analysis family of research typically treats golfing abilities as independent variables while the performance assessment family of research attempts to measure golfing abilities for their own sake. However, both types of research inherently must decide, whether explicitly or implicitly, how golfing ability should be measured. As such, both categories of research are included in this section.

At least as early as 1986, scholars were measuring golfing ability using observed variables. Many of these early studies used separate observed variables as proxies for separate golfing abilities, using them as predictor variables with the goal of understanding which individual abilities mattered the most in terms of certain outcomes such as scoring average or earnings. In 1991, Nix and Koslow (1991) found that they were able to explain 87% of the variance in average round scores and 50% of the variance in prize money earned using four observed variables: average driving distance, the proportion of greens hit in regulation (GIR), the number of putts per round, and sand save percentage.

This style of research has been very popular among those studying golfing ability. Moy and Liaw (1998) used a similar approach, but they focused more on the relative

predictive importances of these variables for PGA Tour golfers. To measure driving distance, they used the average number of yards that a player's drives travel throughout the season. To measure driving accuracy, they used the percentage of drives that land on the fairway on par 4 holes and par 5 holes. They used GIR as a measure of approach shot quality. To measure putting, they used the average number of putts a player uses per GIR for the PGA and Senior PGA tours. For the LPGA tour, they used the average number of putts per 18 holes (presumably due to data availability limitations). Finally, they used the percentage of the time that a player gets the ball into the hole in two shots or less from a greenside bunker to measure ability from the sand. They found that all but one of these abilities have statistically significant relationships with a player's prize money earnings in a given year on the PGA Tour. The one exception was ability from the sand, which did not achieve statistical significance. On the Senior PGA Tour, all of the variables were statistically significant, while, on the LPGA Tour, ability from the sand was statistically significant, but the two driving-related variables (driving distance and driving accuracy) were not.

Other studies have found similar results. Perhaps the earliest example of this type of research, Davidson and Templin (1986) found almost identical results to Moy and Liaw (1998) and Nix and Koslow (1991): GIR, putting ability, and driving distance are all important predictors of success (in that order), while ability from the sand is not. They found that those predictors could explain 86% of the variance in player results, which is very similar to the 87% found by Nix and Koslow (1991). Sharma and Reilly (2013) used a different dependent variable (percentage of tournaments in which a player finishes in

the top 10) and found the same rank order of importance: GIR, then putting ability, then driving distance. Driving accuracy was not a useful predictor.

Shmanske (1992) found that putting ability, driving distance, and approach shot quality are all important factors, and they go in that order: putting ability is more important than driving distance, which is more important than approach shot quality. He found that driving accuracy and ability from the sand were not statistically significant predictors of earnings. Fried et al. (2004) similarly found the same ranked order of skills: putting ability is most important, followed by driving distance, followed by approach shot ability (GIR). They also found that that relationship does not hold for the Senior PGA Tour or for the LPGA Tour. Callan and Thomas (2007) also found that putting is the most important skill for predicting a player's earnings; they found that driving distance, driving accuracy, GIR, and ability from the sand all matter as well. Stemen (2002) found that GIR is relatively more important than driving accuracy, which is relatively more important than driving distance. Rinehart (2009) found that not only are GIR, putting ability, and ability from the sand the most important predictors of earnings, but they are the only relevant ones: driving distance and driving accuracy have no significant effect on earnings for PGA Tour golfers. Focusing on a much smaller professional tour, Botha et al. (2021) found that GIR and putting ability are consistently more important skills than driving accuracy at all quantiles of the ability distribution on South Africa's Sunshine Tour. However, because this tour does not track driving distance, they were unable to compare driving distance to GIR and putting ability. In addition to more commonly used measures of ability like driving distance and putting ability, Finley and

Halsey (2004) found evidence supporting the importance of less commonly used observed variables—scrambling ability and bounce-backs—both of which may inadvertently be measuring some aspect of mental fortitude.

Watkins (2008) found that driving distance and driving accuracy have negligible marginal effects on earnings across the board. In this study, putting ability, GIR, and skill around the green (chipping and pitching) were found to have the greatest marginal returns to skill. Kahane (2010) found similar results, although GIR was at least as important as putting ability in this study. The three most important skills were GIR, putting ability, and ability from the sand. He found that driving distance does have a statistically significant relationship with earnings, but its effect is much smaller than the other three. He also found that driving accuracy is not a relevant predictor of earnings: it was statistically nonsignificant and had the opposite sign than expected. Interestingly, he found that the marginal return to driving distance was higher for players of lower overall ability, while the marginal return to GIR and putting ability was higher for players of higher overall ability. This implies a differential recommendation for improvement to players, rather than a universal one: the area on which to focus most for improvement depends upon one's existing ability level. Rinehart (2009) found an identical rank order of skill importance and also found that the relative importance of these skills has not shifted over time.

Conversely, other studies have found shifts over time in the relative importance of these observed skills. For example, Heiny (2008) found that driving accuracy used to matter (roughly before 2000) but it does not have a statistically significant impact on

earnings anymore. He also found that GIR, putting, and ability around the green (chipping and pitching) are the most important skills, while driving distance also matters. Another study looked at changes in the relative importance of abilities over time and found that driving distance replaced putting ability as the most important ability starting in 2011 (Baugher et al., 2016). Engelhardt (1995) used total driving (a combination of driving distance and driving accuracy) to show that ability off the tee is more important than GIR and that this effect is increasing over time. Engelhardt (2002) confirmed this finding while also demonstrating that total driving ability as a combined metric is a more effective predictor of performance than either driving accuracy or driving distance separately. Bliss (2021) observed that driving distance ability has increased in recent years, but its relative effect on scoring has not: the effect of being able to drive the ball further than other competitors remains the same.

On the other hand, Alexander and Kern (2005) found that while the relative importance of driving distance has increased over time, it is still less important than putting ability. Wiseman and Chatterjee (2006) also found that putting remains more important than driving distance, but they found that the relative importance of driving distance has decreased over time due to its increasingly negative correlation with driving accuracy: because players who hit the ball further than other players are becoming, on average, less accurate, the potential gains of the increased distance are being offset by the penalties of missing the fairway. Broadie and Ko (2009) similarly found, using a simulation model, that much of the benefit of long drives is offset by the decreased accuracy. They concluded that, especially for players who already have above-average

distance, driving accuracy may be more important. Heiny and Heiny (2012) also found that driving accuracy trumps driving distance in terms of importance.

Focusing on the LPGA Tour instead of the PGA Tour, Chae et al. (2021) found that GIR and putting ability are the most important skills, followed by driving distance and driving accuracy, both of which are still more important than ability from the sand. They then used these variables to predict tournament winners, finding that their artificial neural network performed better than their discriminant model, better than their classification trees model, and better than their logistic regression model. This finding likely has relevance for those trying to use current ability estimates to forecast future performance. Kim and Chae (2021) had similar results, but they specified that putting is relatively more important than GIR on the LPGA Tour. Park and Lee (2012) also focused on the LPGA Tour, finding that GIR was the most important predictor of success, followed by putting ability, followed by driving distance, driving accuracy, and ability from the sand, respectively.

Other studies have found differing relative importance rankings among golfing skills depending on the tour. Moy and Liaw (1998) found that ability from the sand matters on the senior PGA tour, but it does not have any effect on the regular PGA Tour. Similarly, they found that ability from the sand is statistically significant on the LPGA Tour while driving accuracy and driving distance are not. Jiménez and Fierro-Hernández (1999) found that all of the skills that they measured (driving distance, total driving ability, GIR, and ability from the sand) were statistically significantly different for the top-ranked players and the bottom-ranked players on the European Tour. Conversely, a

different study applied the same methodology to the PGA Tour and found that only driving distance and total driving ability were statistically significantly different between the top-ranked players and the bottom-ranked players (Engelhardt, 1997). Engelhardt (1999) suggests that these differences imply that the PGA Tour is likely more competitive and has more parity than the European Tour does. Manwaring (2016) found that driving distance is more important on the PGA Tour, approach shot ability (GIR) is more important on the European Tour, and that putting ability is important on both tours.

Unsurprisingly, a player's earnings are strongly affected by the player's performance in the most prestigious events such as major championships and playoff events (Ohn et al., 2012). However, this is somewhat endogenous: the biggest and most prestigious events (i.e. the major championships) also offer the greatest prize money, so it should not be surprising to see success in these events correlated strongly with annual prize money. Ohn et al. (2012) also found that putting ability is a better predictor of earnings than GIR, which is a better predictor of earnings than driving distance, though all three are have statistically significant effects. Similarly, Hamel et al. (2016) found that putting ability is the most consistent predictor of earnings and scoring average. Perhaps more interestingly, Hamel used these results to provide evidence that the PGA Tour has significant barriers to entry and that there are players on other tours who would fare well on the PGA Tour if they were not being excluded from participating by entry rules and requirements.

Most of these early studies using observed variables have found that putting and/or GIR are the most important skills across professional golf tours, at least before

2011. After 2011, the results are more mixed, with driving distance possibly becoming more important. However, these are not universal findings. Dorsel and Rotunda (2001) found that driving accuracy was the most important predictor of success, while driving distance, GIR, and putting are all relevant, statistically significant predictors as well. Belkin et al. (1994) found the same thing, and they additionally found that these results are stable over time. Shmanske (2008) found that using tournament-level data rather than season-level data improved the predictive power of the model significantly. In this model, he also found that driving distance was the most important skill for predicting earnings, followed by putting ability, GIR, and driving accuracy (in that order).

Other researchers have focused specifically on putting. Rather than treating putting as a single ability, they have tried to break it up into separate abilities. Karlsen and Nilsson (2008) did so by measuring green reading ability and technique separately. They found that 60% of the variability in the distance remaining to the hole after a putt can be attributed to green reading ability, 34% to technique, and 6% to green inconsistencies (i.e. random error). Bouvet (2011) focused, instead, on putts of different lengths. Bouvet found that short putts (those under five feet) and long putts (those over 25 feet) were both stronger predictors of success than the medium-length putts between five and 25 feet. Bouvet also found that, in addition to the different lengths of putts, driving distance positively affects performance while driving accuracy has no effect.

In addition to those using separate measures for various (assumedly) distinct abilities, some studies have used unidimensional measures of golfing ability. Baker et al. (2006) used annual scoring average itself as a measure of skill rather than as a dependent

variable to be predicted by other skills. Clark III (2006) found a noticeably weak ($r = 0.17$) correlation between winning match play matches and world ranking, implicitly using world ranking position as a measure of ability, and Coate and Toomey (2014) used a player's position on the money list as a proxy for skill in the context of assessing the performance of caddies. Sen (2012) took a different approach, creating his own measure of composite golfing ability using arithmetic and some logical assumptions. His measure takes the form of the equation: $ability = \frac{\frac{birdie\,\%\,when\,GIR}{failure\,to\,scramble\,\%}}{\frac{1-GIR\,\%}{GIR\,\%}}$, which can be simplified to a slightly more mathematically approachable (though perhaps less intuitive from a golfing perspective) form: $ability = \frac{(birdie\,\%\,when\,GIR)(GIR\,\%)}{(failure\,to\,scramble\,\%)(1-GIR\,\%)}$. He found that this measure is more strongly correlated with a player's annual earnings than any other single observed measure of golfing ability. However, this ignores the possibility that the other measures combined might still be better predictors (through a multiple linear regression, for example). Hoegh (2011) created a measure called the 'performance coefficient' that attempts to measure a player's performance relative to his or her potential using simulation methods.

The Movement to Strokes Gained

In the early 2000s, some golf researchers began to question the ways that they had previously been measuring golfing ability. The 'old' way of using basic observed, season-level descriptive statistics as assumed measures of ability still continues, but the critiques have led to something of a modernization (perhaps even a 'revolution') of golf-related measurement. These critiques began in 2002 with Ketzcher and Ringrose (2002)

arguing that previous academic measurements of golfing ability had incorrectly been treating different observed variables as independent of each other. For example, they pointed out that GIR as an observed proportion can be attributed to skill with driving distance (it is easier for a player to hit the green if he is closer to it), driving accuracy (it is easier for a player to hit the green if he is approaching it from the fairway), and approach shot ability (players will be more likely to hit the green if they have greater ability at approaching the green).[2] Thus, the common usage of GIR as solely a representation of approach shot ability may be misleading. Similarly, scholars had previously assumed that the average number of putts per round represented putting ability. However, Ketzcher and Ringrose (2002, p. 218) point out that the number of putts is likely to also be affected by a player's proximity to the hole rather than just his putting ability: a player making it into the hole in two putts from 72 feet has surely demonstrated greater putting skill than a player making it into the hole in two putts from 11 feet, though the naïve measure of total putts per round will obscure this.

They suggest, therefore, that measures such as the average number of putts per round or the average number of putts per green in regulation actually measure multiple skills simultaneously. As such, these measures might be acceptable as some composite measure of multiple skills, but the naming conventions of previous studies have seemed to imply that each of these is a measure of a specific ability. This is conceptually

[2] Because this study focuses on the most elite male professional golf tours, I proceed from this point by using traditionally male (he/him/his) pronouns when referring to a generic golfer. However, it is worth noting that there are a few women that have also played in male events, and they are included in the dataset discussed in Chapter Three. For example, Lexi Thompson played two rounds at the 2023 Shriners Children's Open in Las Vegas. Although she did not make the cut to proceed to the final two rounds, she did finish with a better score than 33 of the men and tied with 11 others.

problematic: if we are unclear about what skills our measures are actually measuring, it becomes difficult to draw coherent and accurate conclusions (Ketzcher & Ringrose, 2002, p. 218).

Importantly, Ketzcher and Ringrose (2002) did not conclude that there is an easy solution to this problem. Indeed, they argue that it is quite difficult to disentangle existing measures of playing style and abilities. Nonetheless, despite the difficulties in doing so, they argue that scholars of golf research must move in this direction if they want to draw conclusions about the relative importance of separate skills or if they want to use measures of skill and playing style for prediction. Their first movement towards addressing these issues was to construct a model with a random intercept for the tournament, implying that different tournaments have different difficulties. This potentially reduces bias in the estimated relationships between measures of ability and outcomes of interest (earnings, scoring, etc.). This was the first tepid movement among scholars studying golfing ability toward mixed effects modeling.

James (2007) similarly critiqued the literature on measuring golfing ability/performance, suggesting that simple statistics like GIR are not sufficient. He argued that we should only use measures that isolate individual skills rather than ones that represent a composite of multiple skills (James, 2009). However, his proposed alternatives still rely heavily on observed variables, so they do not represent particularly satisfying solutions. One possible exception is the proposal to use approach shot accuracy rather than GIR as a measure of approach shot ability (James & Rees, 2008).

Broadie (2008) further critiqued the existing measures of golfing ability in much the same way that Ketzcher and Ringrose (2002) did, arguing that existing measures have at least two limitations. First, he pointed out that most existing statistics measuring golfing ability actually measure a combination of different skills rather than a single skill. Second, many existing statistics involve proportions and do not characterize the extent of a mistake. For example, if a player misses the green in regulation, simply adding that to the denominator of the GIR calculation does not identify *by how much* the player missed the green. Surely, missing the green by four inches shows more skill than missing the green by 40 yards, but the GIR statistics cannot account for this. Unlike previous critiques, however, Broadie (2008) proposed a more radical solution. He proposed isolating each shot as the unit of analysis rather than the hole, round, or season. By looking at each individual shot, one could estimate how much that particular shot improved or hurt a player's scoring ability on the hole.

Although Broadie (2008) did not yet use this term, this was the very beginning of the movement towards Strokes Gained. Since its creation, Strokes Gained as a metric has revolutionized the way that golfing ability is measured. It began as a book chapter, but the idea has spread in the academic literature; it is now so mainstream that it is reported on the PGA Tour's website, and television commentators reference it casually with the assumption that audiences and players know what it means. The basic idea is that, before a shot, a player has an expected number of shots to get the ball into the hole (based on distance to the hole, playing surface, etc.). After the shot, this number can be recalculated and used to calculate how effective the shot was. For example, imagine that, at a certain

point on the course, a shot has an expected strokes remaining value of 4.2 strokes before the shot. Then the player hits the shot, and the expected strokes remaining is now 3.0 strokes. The golfer used one stroke, but his/her position is now 1.2 strokes better than before. Thus, the golfer "gained" 0.2 strokes with that shot. These strokes gained, importantly, can be summed across a round, tournament, or season.

Of course, this method requires detailed shot-by-shot data that is not always going to be available, especially at lower levels. It is also still inherently based on observed variables rather than latent measures of ability. Nonetheless, it represented (and still represents) an important movement towards disentangling the different dimensions of golfing ability and measuring golfing abilities in an unbiased manner. Broadie's (2008) conclusions were relatively muted in this study, but he used this new measure to show that previous studies that had putting as the most important component of golfing ability may have been biased and misleading. He found that driving distance is easily the most important skill on the PGA Tour.

Fearing et al. (2011) were perhaps even more critical of the state of golf research than Broadie (2008) and Ketzcher and Ringrose (2002) were, arguing that the existing statistical analyses of golf were not useful and that no existing method was able to quantitatively measure how good or bad a particular shot was. They listed three particular drawbacks from which golf research suffered. First, they pointed out that the PGA Tour (and other professional golf tours) only reported a limited number of aggregate statistics, so fans and researchers are forced to rely on them. Second, they pointed out the same issue that other researchers have: the statistics that are reported by the PGA Tour often

are produced by multiple distinct skills and are ineffective for disentangling the different abilities that golfers may possess. Third, they argued that the aggregate statistics are strongly biased by the difficulty of the courses played, since players all choose to play different sets of tournaments/courses throughout a season.

Fearing et al. (2011) then explicitly built on Broadie (2008) to create a statistical method of quantifying the value of individual shots. Instead of all shots, however, they focused specifically on putts. Using a Markov chain statistical technique, they combine logistic regression for the probability of making a putt with gamma regression for the distance remaining after the putt. They also included dummy variables for the hole and for the player, which is another small movement towards the mixed effects modeling framework. The result of their efforts is a metric that they call "putts gained per round," a measure of putting ability that eliminates the bias and confounding present in prior measures of putting ability (Fearing et al., 2011, p. 7). Although it is still based directly on an observed variable framework rather than on a latent variable framework, it represented (like Strokes Gained more broadly) a monumental improvement in the measurement of golfing abilities. They also showed that this is not just an academic exercise: the differences produced between the new measure and the more traditional measures of putting ability are notable. For example, they point out that Vijay Singh ranked 8th according to "putting average" from 2003-2008, but he ranked 218th in the new Putts Gained metric (Fearing et al., 2011, p. 34). This is likely due to Singh possessing some superior skills other than putting: he was getting the ball so close to the hole that he

didn't use very many putts, but, once putting ability alone is isolated, he does not appear to be a particularly adept putter (relative to others on the PGA Tour).

In 2012, Broadie built upon his previous research to formalize the Strokes Gained formula (Broadie, 2012). This article introduced Strokes Gained formally for the first time (since he did not use that term in his prior research) and presented the math behind its calculations. To determine the exact average value of each shot from a given distance and surface, he used polynomial regression with course-round random effects to represent the difficulty of each round played. It was at this point that Strokes Gained began to receive widespread recognition and use beyond academia. It quickly became the state-of-the-art method for assessing player abilities on the PGA Tour. Unfortunately, the PGA Tour is the only professional golf tour to track and provide shot-level data, so this measure has still only gained widespread adoption on the PGA Tour. Other tours still rely on more traditional statistics.

Additionally, it is worth pointing out that Stroked Gained still has some limitations. Most importantly, even if Strokes Gained has partially solved the degrees of freedom problem of multiple abilities being measured at the same time, it is still possible for a player's ability in one domain of golfing ability to influence his or her Stroked Gained values in another. For example, imagine a player that is so good in terms of driving ability and approach shot ability that he hits the ball to 2 inches from the hole every single time. Because every player would be expected to make a 2-inch putt every single time, his Putts Gained will be 0, implying that he is exactly average at putting. In reality, though, we have no idea what his putting ability is: he could actually be well

above average or well below average, and we would have no way of knowing. The true ability is entirely attenuated toward zero in the measure of golfing ability. This, of course, is an extreme example, but it shows a potential weakness of Strokes Gained. In practice, of course, nobody is going to hit the ball to two inches every time, so our estimates would never be this ambiguous. Nonetheless, because Strokes Gained is additive, it is a real issue. Broadie (2012) found that the correlations between different skills (e.g. putting, driving, approach shots) using the Strokes Gained approach are low, which may provide at least some evidence that this problem is not prohibitive or insurmountable.

A second weakness of Strokes Gained as it is typically implemented is that it implicitly assumes a factor structure. For example, the PGA Tour provides data on Strokes Gained total, Strokes Gained Off-the-Tee, Strokes Gained Around-the-Green, Strokes Gained Tee-to-Green, Strokes Gained Approaching the Green, and Strokes Gained Putting. These are logical ways to divide individual golf shots, but such divisions are still made entirely by assumption. No rigorous empirical analyses have been employed to demonstrate that these are distinct skills, that they are the only distinct skills, or that they provide a good fit to the shot-level data. Thus, anyone who uses these Shots Gained calculations (other than the "total" category) is accepting an untested assumption.

A few small variations to Strokes Gained have arisen in recent years. Heiny and Heiny (2014) followed the same logic as Strokes Gained, but they used a Markov chain rather than polynomial regression to model the number of strokes that is expected from specific distances and locations. Similarly, Chimka and Talafuse (2016) used Poisson regression instead of polynomial splines to model the number of strokes that is expected

from specific distances and locations. Drappi and Co Ting Keh (2019) used machine learning instead of the polynomial splines to predict the remaining number of strokes to complete the hole, and they introduced Smoothed Strokes Gained, which adds a penalty to the Strokes Gained formula for long hole-outs due to the likelihood that long hole-outs are due in part to luck rather than skill.

Drappi and Co Ting Keh (2019) also provided perhaps the clearest argument against the prior generation of research that used individual observed variables as proxies for different abilities. They argued that these older studies made a very reasonable assumption that traditional PGA golfer aggregate skills could be used as measures of ability and then used to predict performance. However, Drappi & Co Ting Keh (2019) also argued that the inconsistencies in the findings from these studies point to some underlying problem with this assumption: "one has to question the variability of the coefficients of the regressions and the variability in which skill factors are most important across these studies" (Drappi & Co Ting Keh, 2019, p. 66). Their explanation was that the highly correlated nature of the observed aggregate statistics and their overlapping nature produced unstable results. Thus, Strokes Gained represents a huge improvement. Drappi and Co Ting Keh (2019) confirm this outright, crediting Broadie's (2008; 2012) Strokes Gained metric as a massive improvement over the previous generation of golf measurement studies.

Perhaps the most significant proposed alteration or improvement to the Strokes Gained measure has been the proposal of the ISOPAR Method for measuring an individual golf shot's value. It relies on the same logic as Strokes Gained—that the

difference between the expected number of strokes before a shot and after the shot can represent the quality of that shot. However, whereas the Strokes Gained metric attempts to generalize (or average) the expected number of strokes remaining given a certain distance and a certain surface across holes and courses by controlling for the effects of all relevant factors statistically in a regression setting, the ISOPAR Method requires and creates a unique map for each hole with bars similar to a map of barometric pressure or to a topographical map. Thus, for a particular hole, the ISOPAR method can capture whether it is better to miss the fairway to the left or the right, whether it is better to miss the green in the sand trap or in the rough, etc. This does not need to be generalized to all holes; it can be unique to that specific hole. This idea was first introduced by Stöckl et al. (2011) and Stöckl et al. (2012). From a conceptual perspective, it is an appealing improvement to Strokes Gained that very much follows the same logic. However, from a practical perspective, it would require large amounts of shot-level location data from every shot played on every hole on every golf course to create the maps in the first place, and it would require similarly detailed data to then calculate a shot value for each shot for each player and then aggregate these to calculate an ability value. That goes beyond Strokes Gained in its data collection requirements, which are already onerous enough that only one professional tour tracks it. This may be why the ISOPAR Method has not gained much traction. Even its creators have mostly applied it only on the greens rather than for the entire course (e.g., Stöckl et al., 2011; Lamb et al., 2011).

Since the creation and popularization of the Strokes Gained metric, several studies have used players' Strokes Gained values as explanatory variables. Heiny and

Frisby (2018) used Strokes Gained values in an ordinal logistic regression to predict players' scores on individual holes. Aparicio et al. (2021) used Strokes Gained values in a survival model to predict how long it will take before a given player on the PGA Tour wins his first tournament. They found that power/driving distance is the most important factor, followed by putting ability, followed by approach shot ability. Korpimies (2020) used season-level Strokes Gained averages to predict players' performance in future seasons and evaluate which prediction model is most effective. The results showed that the random forest model performed better than the logistic regression model.

Due to the relatively recent popularization of Strokes Gained, relatively fewer studies have used these metrics as predictor variables thus far compared to the number of studies that have utilized counting statistics. Among those that have, however, one noteworthy feature is that they consistently show that power/driving distance is the most important skill. Theoretically, this may be due to changes in the game of golf. However, it is likely that this improved consistency is due at least in part to the decrease in measurement error and multicollinearity that has resulted from the shift from aggregated season-level observed statistics to Strokes Gained. Nonetheless, the more traditional metrics are still commonly used, especially in the context of smaller tours that do not provide or calculate Strokes Gained.

Golfing Ability as Latent Variable(s)

Classical Test Theory Approaches

Although it was not directly framed this way, the first effort to treat golfing ability of professional golfers as a latent variable was published in 1995. Making an analogy

between items on a test and golf holes, Clarke et al. (1995) used the methods of Classical Test Theory to assess the discrimination of individual golf holes and the reliability of 18-hole rounds of golf and 72-hole tournaments of golf. They found that individual holes were not very effective at discriminating between better and worse golfers, at least with the population restricted to those who have already achieved the status of being professional golfers. Given this result, it is not surprising that they also found that 18-hole rounds of golf and 72-hole tournaments were not very reliable measures of golfing ability. From a viewership and marketing perspective, this may actually be desirable—if fans already knew which player was going to win ahead of time, they may not bother watching. However, from a measurement perspective, this is less desirable, as it means that a relatively large proportion of the variability in players' scores on a given day can be attributed to random variation rather than to true golfing ability.

Indeed, Clark III et al. (2008) also applied Classical Test Theory logic to individual golf holes, and they confirmed the finding from Clarke et al. (1995) that individual holes are not very effective at discriminating between the abilities of professional golfers. Interestingly, though, they found that the opposite was true for highly-skilled amateur golfers and club professionals (individuals who work as golf instructors at golf courses rather than competing on a professional tour): golf holes do effectively discriminate between these players. Thus, the lack of capability to discern players' abilities on the professional tours was likely due to a restriction in range rather than an actual lack of capability for discrimination. The professional golfers on the PGA Tour are *so* good, and their ability levels so similar (compared to the entire population of

golfers), that scores on individual holes are due more to random variation than variation in golfing ability. However, once the range of ability level is widened somewhat (though not entirely, as the authors still focused on golfers with well-above-average ability levels), then scores on individual holes become more useful as indicators of a golfer's ability level.

Among the early scholars to treat golfing ability as a latent variable, Sachau et al. (2009) perhaps most coherently made the case for why this is necessary and most explicitly adopted the language and techniques of Classical Test Theory. They used an example from the 1974 U.S. Open tournament. The course was set up to be so hard that week that that event is known as the 'massacre at Winged Foot.'[3] One of the players, Hale Irwin, commented on the difficulty of the course that year by suggesting that the USGA (the organization that runs the tournament) was "trying to embarrass the best players in the world." USGA official Frank Tatum responded directly: "we had no intention of confounding the best players in the world. We simply wanted to identify who they were." This is the exact logic of a test from the perspective of educational measurement—if the test is too easy (or, less commonly, too hard), then it will not be very effective at identifying the ability of interest (i.e., discriminating). Thus, Sachau et al. (2009, p. 52) argue that "Mr. Tatum's retort illustrates the USGA's view that the purpose of a tournament is to sort the greatest players from the not quite as great. In this regard, the U.S. Open is a type of *test*." By explicitly treating a golf tournament as a test,

[3] Winged Foot is the name of the course on which the tournament was played that year. It is located just outside of New York City.

they were able to utilize some of the typical statistical techniques that are regularly applied to academic tests.

In particular, Sachau et al. (2009) utilized Classical Test Theory to estimate the validity of each course/tournament as a test of a golfer's ability. They used the golfer's mean score over a large number of golf rounds as an analog to the "true score" in Classical Test Theory (Sachau et al., 2009, p. 55). They then used the score during a particular tournament as the test performance measure. Then, they calculated the correlation between the test performance measure and the "true score." This value then represented the validity coefficient: "just as test developers typically reduce measures of criterion-related validity to a simple correlation between a test score and a measure of competence, the validity of a tournament can be measured by the correlation between PGA scoring average and tournament score" (Sachau et al., 2009, p. 55). They then performed a linear regression for each tournament, using tournament scores to predict "true" scores. The slope of this line is equivalent to the discrimination parameter in educational measurement, and the intercept is equivalent to the difficulty parameter. Among the regressions, they found strong positive correlations between intercept, slope, and validity: harder courses do a better job of identifying golfing ability than do easier courses/tournaments.

In practice, Sachau et al. (2009) might have been better off using a single mixed effects model rather than a separate regression for each tournament. Nonetheless, their methods and results are compelling. Perhaps more important than the results themselves, Sachau et al. (2009) explicitly (and successfully) modeled a golf tournament as a test and

considered golfing ability to be a latent variable that each test (golf tournament) was attempting to measure.

Advanced Latent Variable Approaches

Fisher (1998) became the first to move beyond Classical Test Theory into the more modern logic of true latent variable models. He applied a many-facet Rasch model that simultaneously estimated player ability, the relative difficulty of each hole, and the relative difficulty of each of the four rounds to the 1990 US Open tournament. Because the model used was a variant of a Rasch model, each hole is assumed, mathematically, to have the same discrimination value—they each provide an equal amount of information about golfing ability. He found that the players that have the greatest ability according to the model were also the ones near the top of the leaderboard in the tournament, indicating some validity for the measurements coming from the model. Fisher's (1998) conclusion was mostly just about feasibility: golfing ability and golf scores can, in fact, be effectively modeled using latent variable models that are common in the field of educational measurement.

A number of other scholars have moved less explicitly towards treating golfing ability as a latent variable (and, inherently, each hole or round as a test of that ability). Connolly and Rendleman Jr (2008) used a crossed mixed effects model to measure player ability and course-round difficulty simultaneously. Essentially, they used round scores as the dependent variable with separate random effects for the player and for the course-round. The values from the random effects then represented player ability and course-round difficulty, respectively. Interestingly, the authors were most interested in whether

there was evidence of streaky play from round to round within players (positive autocorrelation); they found that there was such evidence. They also found an error term with standard deviation of 2.69, which is quite large. This means that even after accounting for all player skill and course difficulty, a given score would vary on average by 2.69 strokes from the "true" score, or average predicted score. Phrased differently, to encompass 95% of the possible score outcomes for a given player on a given day, one would need a margin of error of 5.38 strokes. If that player's "true" expected score for that day (based on his ability level and the course's difficulty) for the given round is 69.5, we would be 95% confident that he would score between 64.12 and 74.88. That is a very wide range, and it demonstrates that there is a lot of random variation in individual golf scores. They are not particularly reliable measures of golfing ability. Berry (2001) presented a very similar model, with random effects for player ability and course difficulty. He found a random error term with a standard deviation of 3.12, which is even larger than the one found by Connolly and Rendleman Jr (2008). This is likely due to the use of a course-level random effect rather than the round-level course-round random effect used by Connolly and Rendleman Jr (2008): the additional variation in scores that could be explained by measuring differences in the difficulties in individual days/rounds on the same course is moved to the error term. Berry (2001) also found that the standard deviation of the intrinsic abilities of golfers was 2.10 in his model, indicating that the

effect of random variation often has a larger effect on a player's score than his ability, at least in this particular model.[4]

Shmanske (2009) used dummy variables for course difficulty and player ability, which is a slightly less sophisticated modeling technique than using a true mixed effects model. Nonetheless, the logic is similar: the dummy variables for each course create a single estimate of the course's effect on scoring (i.e., its difficulty) while the dummy variables for each player estimate the player's effect on scoring (i.e., the player's ability). In their early movement towards Strokes Gained as an observed-variable approach to measuring golfing ability (mostly just putting ability), Fearing et al. (2011) used a similar approach at the hole-level: dummy variables for each hole and for each player. Pope and Schweitzer (2011) use this same hole-level dummy variable approach while measuring loss aversion among professional golfers. In her efforts to measure the incentive effects of being paired with specific playing partners, Brown (2011) similarly used dummy variables for each player-course combination.

More statistically sophisticated mixed effects models (with random effects rather than dummy variables) have become more common in recent years. In his formalization of Strokes Gained, Broadie (2012) used a course-round random effect to account for the

[4] Note that, in many measurement models, standardized residuals and variance components are the norm. This is at least partially due to the lack of any naturally meaningful scale for the latent variables that they represent. However, most measurement models that measure golfing ability are left in a raw, or unstandardized, form. This is because there is a naturally interpretable and meaningful scale for golfing ability, which is the number of strokes. Thus, we can interpret the residual error term as the average number of shots that an observed score will deviate from the score predicted by the model, and we can use a random effect's standard deviation to calculate how many shots apart from each other randomly selected players are, etc. I follow this convention of leaving the random effects and residuals in unstandardized form as well in later chapters.

varying difficulties across golf courses and the varying difficulties across days/rounds at each course. Although Heiny and Heiny (2012b) were operating primarily from an observed variable perspective, they did add a player-level random effect to their model. Because they also included observed variables for driving distance and driving accuracy, they interpreted the player-level random effect as representing a player's ability outside of driving distance and driving accuracy. Heiny and Heiny (2012a) followed the same logic; because they already measured and accounted for total driving ability using observed variables, they interpreted the player-level random effect as representing an amalgamation of the player's other skills, such as putting ability, approach shot ability, etc. Similarly, Heiny and Frisby (2018) used a player-level random effect as a measure of how well a given player performs at the Master's tournament specifically (an affinity measure, in other words), since they were already accounting for player ability using Strokes Gained components as predictor variables. In the context of trying to discern whether streakiness, 'hot hands,' and 'cold hands' exist in golf, Elmore and Urbaczewski (2018) also implemented player-level random effects to denote player ability.

Connolly and Rendleman Jr (2012) implemented perhaps the most complicated and comprehensive version of a mixed effects model of golfing ability. They included round-course random effects to capture the difficulty of that particular day of golf,[5] they implemented player-course random effects to capture a player's affinity with a particular course, they implemented time-varying player-level random effects, and (naturally) they

[5] Note that this would be equivalent to a random effect simply for the round if every tournament were played on a single course. However, there are a few tournaments that are played on multiple courses, so specifying that it is the course-round rather than just the round is relevant.

had an error term that they interpreted as luck. These time-varying effects not only measured player ability, but they also allowed that ability to change over the course of time. Since they used a relatively long time period in their sample (2003-2009), it makes sense that an individual's ability level could change over this period, so the time-varying random effects are an interesting addition. In a related study, the authors employed the same model to show that reasonable deviations in the current (as of 2012) format of the year-end playoff format known as the FedEx Cup would not yield any improved efficiency in terms of rewarding better players (Connolly & Rendleman Jr, 2012).

Broadie and Rendleman Jr (2013) provided the clearest explanation for how a mixed effects model can work as a latent variable measurement model for golfing ability. They called this their "Score-based Skill Estimate," or SBSE (Broadie & Rendleman Jr, 2013, p. 130). They described this estimate as providing "an estimate of [a player's] mean 18-hole score played on a 'neutral' course in which the common effects of round-to-round variation in scoring due to differences in intrinsic course difficulty, course setup, weather, etc. have been (statistically) removed" (Broadie & Rendleman Jr, 2013, p. 130). Thus, any difference between players' estimated ability levels predicts the mean difference in their scores on a given course on a given day. As such, this estimate can very clearly be interpreted as an estimate of a player's golfing ability. One key benefit of this method is that it "allows golfers who never play on the same course to be ranked on a single scale as long as there are other golfers to link them together" (Broadie & Rendleman Jr, 2013, p. 131). This is analogous to the scoring of tests in educational measurement: two individuals can take different tests (or different forms of a test) and

still receive valid scores on the same scale as each other because the tests either have items linking them (the same item on both tests/forms) or individuals linking them (the same individual taking both tests/forms). In this case, it is the latter. Because of this benefit, the model can incorporate players from multiple tours together into a single rating system.

Although the Official World Golf Rankings (OWGR) attempt to rank players from different tours as well, Broadie and Rendleman Jr (2013) show that these rankings are biased. Specifically, they are generally biased against the players on the PGA Tour (the most prestigious and elite tour in the world) by an average of 26-37 spots—a statistically significant difference. This effect is strongest, they found, for players with SBSE rankings between 40 and 120. Thus, using a latent variable method such as SBSE is likely a more effective way to measure golfing ability than simply looking at the OWGR. Indeed, it may not be surprising that the OWGR system is biased, as its creators have an incentive to encourage players to play more tournaments, as this drives revenue. As such, they may reward players who play more tournaments with more points and subsequently higher rankings. However, the number of tournaments played by a given player should, statistically, be unrelated to his ability level.

The model presented by Broadie and Rendleman Jr (2013) is relatively parsimonious: $S_{ij} = \mu_i + \delta_j + \varepsilon_{ij}$, where S_{ij} is a player's score in a given round, μ_i is a player's ability level, δ_j is the course-round difficulty, and ε_{ij} is the error term (random variation). They provided the clearest argument for and explanation of this type of model. However, they explicitly cite Broadie (2012), Connolly and Rendleman Jr (2008), and

Connolly and Rendleman Jr (2011) as previous examples/originators of this type of model. Stigler and Stigler (2018) later implemented an even simpler version of this model, eliminating the course-round random effect so that the intercept in the model represented mean player ability and the player-level random effect represented how far above or below average that player is compared to that intercept.

Methods Literature

Mixed Effects Modeling

What are Mixed Effects Models?

Mixed effects modeling is, perhaps confusingly, known by many different names. The models produced using this statistical technique are sometimes called multilevel models, mixed-effects models, random-effects models, random-coefficient regression models, covariance component models, and hierarchical models (Raudenbush & Bryk, 2002, pp. 5–6). The two most common of these are probably "hierarchical" models and "multilevel" models, as they describe the nature of the data as well as the characteristics of the statistical model. The data on which one might apply a hierarchical model or a multilevel model are structured in a way that has some sort of grouping of observations/individuals within broader categories of observations/individuals. Education is a classic example of this: students are nested within teachers, who are nested within schools, which are nested within districts, which are nested within states, etc. Similarly, college graduates may be grouped or nested within their universities or within their majors (or both). Nesting or grouping can occur in many other settings as well. For example, country-year observations in political science are probably nested within the

broader category of country, hospital patients may be nested within hospitals and/or within regions, and individual dogs may be nested within breeds.

Regardless of name and context, the format of mixed effects models is essentially universal: these models "were introduced mainly for modeling responses of individuals that have the same global behavior with individual variations" (Kuhn & Lavielle, 2005, p. 1020). Using these models may be useful for testing specific hypotheses about cross-level effects and for partitioning variance among the different levels/categories (Raudenbush & Bryk, 2002, p. 7). For example, one could estimate the effect of the teacher on student performance relative to the effect of the school district. In these cases, the researcher's explicit interest in the multilevel structure of the data makes mixed effects modeling a logical choice. However, despite the frequency with which social science research and data have some sort of nesting or hierarchical structures, past studies have very often neglected to address these structures adequately (or at all) in their modeling approaches. At least part of this neglect has been due to the historical difficulty in estimating these models (Raudenbush & Bryk, 2002, p. 5). Fortunately, modern statistical software can handle these models relatively easily, so this is no longer as much of a barrier.

Additionally, even if one is not directly interested in the nesting structures of the data, it is important to account for them in the statistical analysis of data that are generated with a natural nesting or hierarchical structure. Without accounting for the nesting structure, the results of the traditional statistical model—typically a linear regression model or a generalized linear regression model (GLM) such as logistic regression—have the potential to be biased. In many ways, this is similar to the logic of

omitted variable bias: if there is a variable that is related to both the predictor and the outcome, then running a regression (or t-test or ANOVA, etc.) will attribute some of the effects of the omitted variable to the included predictor variable, leading to biased estimates for the true effect of the predictor variable. Similarly, if the hierarchical structure of the data is ignored, then some of the effects that occur due to that structure may be incorrectly attributed to other factors, and the results will therefore be biased (Tuerlinckx et al., 2006, p. 226).

Thus, one of the most important (though sometimes overlooked) uses of mixed effects modeling is to reduce bias in the context of regression analysis. Gelman and Hill (2006, pp. 6–7) call this use of mixed effects modeling "analysis of structured data." This use sounds very generic. It basically means that a researcher may not be directly interested in the structuring/multilevel/nesting mechanisms, but the researcher still uses mixed effects modeling to get the most accurate, unbiased estimates in the presence of (possible) hierarchical structures.

There are two primary ways to conceptualize what a mixed effects model is actually doing. First, one could think of a mixed effects model as a regression model in which "each of the levels in [the] structure is formally represented by its own sub-model" (Raudenbush & Bryk, 2002, pp. 6–7). This takes the form of various coefficients in the primary regression model also being modeled as dependent variables as part of a separate sub-model; the sub-models can then all be combined to create a single comprehensive model (Gelman & Hill, 2006, p. 235). Conceptualizing and writing mixed effects models

in this way is useful from a conceptual or pedagogical perspective, particularly when there is only one grouping variable or when the grouping variables are entirely nested.

However, a second (mathematically equivalent) conceptualization of mixed effects models is often more efficient from a practical perspective. From this perspective, we may think of mixed effects models as "extensions of regression in which data are structured in groups and coefficients can vary by group" (Gelman & Hill, 2006, p. 237). This is where the terminology gets a bit confusing. The coefficients that we allow to vary in the mixed effects model "are sometimes called *random effects,* a term that refers to the randomness in the probability model for the group-level coefficients" (Gelman & Hill, 2006, p. 245). The term "fixed effects" is used in contrast to random effects, but the term unfortunately is not used consistently. In some cases, "fixed effects" refers to the components of a model that do not vary by group: a particular mixed effects model may have some variables that have only "fixed effects" (no random effects) and other variables that have both "fixed effects" and random effects. In other cases, however, the term "fixed effects" may refer to the use of dummy variables for each level/category of some grouping variable. For example, if the authors of a study describe using "country fixed effects," this likely means including a dummy variable for each country (except for one, which is excluded as a reference category to prevent perfect collinearity). Even more confusingly, then, is when researchers take these different parameter-level definitions of "fixed effects" and use them to describe the entire model. Does a "fixed effects model" mean one in which dummy variables for categories were used? Does it mean one in which no coefficient varies by group (i.e., no random effects)?

This confusion leads Gelman and Hill (2006, pp. 245–246) to conclude that we should try to avoid the use of the terms "fixed effects" and "random effects." They much prefer the terms "hierarchical models" and "multilevel models" instead: they end up selecting "multilevel model" as their preferred term, while Raudenbush and Bryk (2002) end up using the term "hierarchical models." I use the terms "multilevel models," "hierarchical models," and "mixed effects models" interchangeably, though I mostly use "mixed effects models" when possible. Thus, at the model level, I mostly follow Gelman and Hill's (2006) lead and avoid the use of "fixed effects" and "random effects." However, at the parameter level, there is sometimes a need to refer to a specific effect/parameter/coefficient and indicate whether or not it varies by group/category. In these cases, I refer to a "random effect" as a parameter that is modeled as varying by group or category. Thus, in an educational context, a "school-level random effect" would mean that each school gets its own value for that particular parameter, modeled as a single draw from a probability distribution. Conversely, I use the term "fixed effect" to mean a parameter that is equal for all values/levels of the grouping variable.

In the regression modeling context, it is common for a variable to have both a fixed effect and a random effect. In this case, the fixed effect represents the average effect of that independent variable on the dependent variable across the different categories or groups, while the random effect represents the deviation from that average effect for a particular group or category This type of model and this interpretation are widely accepted: "the notion that individuals' responses all follow a similar functional form with parameters that vary among individuals seems to be appropriate in many

situations" (Lindstrom & Bates, 1990, p. 673). Similarly, though slightly more formally, Rabe-Hesketh et al. (2004, pp. 167–168) view this in terms of unobserved heterogeneity between groups: random intercepts represent heterogeneity between clusters in the overall level of the dependent variable, while random effects on coefficients represent heterogeneity between clusters in the relationship between the given independent variable and the dependent variable.

Although some research does indeed use dummy variables to represent group or category membership, such a model is never going to perform better (in terms of efficiency, accuracy, etc.) than a true mixed effects model. As such, Gelman and Hill (2006, pp. 245–246) give the advice to *always* use mixed effects modeling, even though the literature gives varying and contradictory advice. In many ways, the two options are similar, with one important difference: in the mixed effects context, the coefficients are themselves modeled—sometimes by a relatively simple shared probability distribution and sometimes by a more complex regression model that incorporates predictors from other levels (Gelman & Hill, 2006, p. 252). Even in its simplest form (where the group-level coefficients are modeled simply as random draws from a probability distribution), this provides two benefits over the dummy variable approach. First, all levels/groups can be modeled, and we do not need to exclude one as the reference category. This eases the interpretation of the group-level values and no longer requires the researcher to select which group to use as the reference category. Second, it allows for borrowing of information between observations/individuals within a group or category and across groups. This improves the efficiency of the estimation, and it allows for models in which

some groups have as few as one or two observations (Gelman & Hill, 2006, p. 276). As such, Gelman and Hill (2006, p. 9) recommend mixed effects modeling in almost all regressions using observational or structured data.

Lee et al. (2020) provide a nice demonstration of the benefits of the information 'borrowing' discussed by Gelman and Hill (2006). They show that a mixed effects growth model grouping outbreaks of COVID-19 by country outperforms existing models that were based on individual countries separately. They directly credit this to the information borrowing capabilities: "because the proposed model takes advantage of borrowing information across multiple countries, it outperforms an existing country-based model" (Lee et al., 2020, p. 1).

In contrast to many in the fields of psychological and educational statistics, Gelman and Hill (2006) do not recommend model or variable selection based on statistical significance, nor are they particularly worried (statistically) about parsimony and over-fitting. Instead, they argue that the main constraints are actually human limitations and software limitations, rather than any statistical or mathematical need for parsimonious models: the main reasons not to use extremely complicated models with random effects on every coefficient are human inability to interpret such complicated models and software inability or inefficiency in estimating such models (Gelman & Hill, 2006, p. 271). In other words, such complex models with many variables, many levels, and many random effects are likely to be statistically valid, even if software may struggle to actually estimate them and humans may struggle to be able to interpret the results. Raudenbush and Bryk (2002, p. 384) are somewhat more skeptical of these complex

models, but this is mostly due to concerns about software estimation and whether the researcher will have sufficient data to estimate them.

Mixed Effects Models as Measurement Models

Mixed effects modeling also forms the basis for most latent variable statistical models, even though many users of these methods may not be aware of this. In effect, the random effects that are utilized in mixed effects modeling end up representing latent variables. Latent variables that are measured in educational statistics are often just random effects in a mixed effects model, and most random effects in mixed effects models can be interpreted as latent variables: "in reality, [the entire field of hierarchical models] may be viewed as dealing with latent variables" (Raudenbush & Bryk, 2002, p. 337).

Indeed, many scholars have pointed out that existing latent variable measurement models are simply reformulations of mixed effects models (or vice versa). Bauer (2003) and Curran (2003) both showed that linear mixed effects models can be respecified as structural equation models (SEMs), which are typically used, at least in part, as latent variable measurement models. In fact, Curran (2003, p. 529) argued that the equivalence between SEM growth models and mixed effects growth models has been known since the 1980's among statisticians. Curran (2003, p. 565) also discusses this in the broader types of models beyond growth models, indicating that latent variable measurement models (the original domain of SEM) can be run as both SEMs and as mixed effects models, and so can regressions with random effects (the original domain of mixed effects modeling) be run in both ways. Thus, we may be approaching the conclusion that SEM and mixed

effects modeling may differ only in their historical origins and the software used to estimate them rather than in the underlying statistical models themselves (Curran, 2003, p. 565). In other words, the differences between these models no longer exist mathematically (if such differences ever existed at all). Rabe-Hesketh et al. (2004) would agree with that assessment of mathematical similarities despite historically different origins, pointing out that the techniques have notable similarities and function similarly, despite being developed in parallel. Most importantly, one of these similarities is the direct inclusion of latent variables in both types of model.

A specific class of frequently-used latent variable measurement models is known as item response theory (IRT) models. Like other latent variable measurement models, they can be respecified as mixed effects models (Rijmen et al., 2003; Van den Noortgate et al., 2003; Adams et al., 1997; Tuerlinckx et al., 2006; Kamata, 2001). Rabe-Hesketh et al. (2004) presented a unifying framework for measurement models and mixed effects models, and they built a package to convert between the two in *Stata*. Rijmen et al. (2003) argued that the transferability between the two classes of models is useful for those wanting to conduct IRT models; as almost any statistical software can run mixed effects models, scholars and practitioners would no longer need any special software or packages to use IRT models. Van den Noortgate et al. (2003) argues that reformulating IRT models as mixed effects models may even be a better option than their traditional formulation, though it would change some of the interpretations of certain parameters.

An important point is that mixed effects models already are, mathematically, equivalent to latent variable measurement models. Thus, the random effects in mixed

effects models are already measuring latent variables, whether we conceptualize them that way or not. By choosing to do so, we allow our mixed effects models to also serve, simultaneously, as measurement models. Indeed, even though it is not necessarily obvious to those using mixed effects models simply to reduce bias, many scholars have pointed out that random effects in the mixed effects modeling framework are equivalent to latent variables in the measurement context. In this vein, Rabe-Hesketh et al. (2004, pp. 167–168) pointed out the equivalence between latent variables and random effects. They also observed that a similar interpretation of random effects is common in biostatistics and biometrical genetics (Rabe-Hesketh et al., 2004, p. 168). Tuerlinckx et al. (2006, p. 226) also followed this equivalence: "the cluster-specific parameters in mixed models are also called *random effects* or, as in the literature on item response theory (IRT), *latent traits* or *latent variables.*" Thus, it is justified and accepted to treat random effects in a mixed effects model as estimates of latent variables.

Types of Mixed Effects Models

Mixed effects models can follow three broad forms: linear mixed effects models, generalized linear mixed effects models, and nonlinear mixed effects models. Linear mixed effects models are similar to linear regression with random effects added to certain parameters to account for grouping effects. Generalized linear mixed effects models are analogous to generalized linear models, where the outcome variable is not continuous (e.g., logistic regression for binary outcomes, negative binomial regression or Poisson regression for count data, etc.), so the assumptions of linear regression (normally distributed residuals, most notably) are likely to be violated. In these models, the linear

predictor ($\beta_0 + \beta_1 * x_1$ ...) is wrapped inside of a mean function to produce predicted values on the scale of the outcome variable.[6] These models are technically nonlinear, but they are transformations of the linear model (i.e. "linear in the parameters"), so they are readily understood by those familiar with linear modeling. As with the linear models, it is easy to conceptualize how to add random effects to particular parameters in these models to convert the models into mixed effects models.

True nonlinear models—those that are not linear in the parameters and can therefore not be represented as simple transformations of a linear predictor—are often (though not exclusively) used in the context of growth models over time. In these nonlinear models, parameters are able to take on special interpretations beyond being simple multiplicative weights/coefficients for variables: they can represent asymptotes, midpoints, vertices, inflection points, etc. Random effects on these parameters allow for individuals to each have their own unique estimates for these values—their own estimated upper or lower asymptote, their own estimated inflection point, etc. Raket (2020) used this type of model—an exponential decay model specifically—to estimate individual cognitive decline over time in the context of Alzheimer's diagnosis. Jonsson et al. (2000) showed that nonlinear mixed effects models more accurately capture the effects of pharmacological drugs at different doses than do linear or generalized linear models, especially at levels of dosage outside the range of those directly tested.[7] Lee et al.

[6] These can be alternatively conceptualized with a link function, which is just the inverse of the mean function—it wraps the outcome variables inside of a function to convert them to a linear format.

[7] Note that this is not a growth model over time, but rather a growth model with respect to the dosage.

(2020) used a mixed effects Richards growth curve model to model the outbreak of COVID-19 cases over time in separate countries. Furthermore, all of the DMM studies mentioned in the next subsection would fall into this category as well.

A final subcategory within mixed effects models is crossed models. Crossed models occur when there is more than one grouping variable, and neither is entirely a subset of the other. In the earlier example of students nested inside of schools, which are themselves nested inside of school districts, the model is *not* crossed: each student exists entirely within one school and each school exists entirely within one school district. Thus, the "nesting" label is appropriate. However, if one is interested in students being grouped within schools and within neighborhoods, there may be some crossover: some students from a neighborhood may attend one school while others from the same neighborhood attend another school. Thus, two students may live in the same neighborhood but attend different schools, or they may live in different neighborhoods but attend the same school. In the context of golf, of course, each round of golf may be nested both within the player who achieved that score and within the course on which it was achieved, etc. Crossed models may also be called "non-nested models" (Gelman & Hill, 2006, p. 244) or "cross-classified models" (Raudenbush & Bryk, 2002). Regardless of terminology, the key point is that units at lower levels can share in one grouping characteristic but differ in another. Visually, both crossed models and pure nested models could be conceptualized using Euler diagrams; in the crossed model, the two shapes would partially overlap, whereas in the pure nesting model, one shape would exist

entirely inside of the other (representing the nature of one set being entirely a subset of another).

Dynamic Measurement Modeling (DMM)

Within the category of nonlinear mixed effects models, dynamic measurement modeling (DMM) has relatively recently emerged as a specific type of educational measurement model. More traditional psychometric models (e.g., item-response theory models) are quite effective at generating estimates of an individual's ability at a single point in time. However, researchers and practitioners are often interested in an individual's capacity for future development of an ability instead of or in addition to the level of ability that the individual has at a fixed timepoint (Sternberg et al., 2002, p. 142). Without a good way to measure this, researchers and practitioners have often made the assumption, whether implicitly or explicitly, that current ability is a proxy for future capacity. Those developing the tests may not intend for the tests' results to be used in this manner, but it nonetheless seems to be true that "student scores from single-administration assessments are often interpreted not only as pertaining to past and present student performance but also as indicators of student potential to learn in the domain being assessed" (Dumas & McNeish, 2018, p. 612).

This assumption, however, can be problematic. The assumption is often untested, and it can yield distributional biases in high-stakes testing environments, which in turn may harm disadvantaged populations. If certain populations have had fewer opportunities to access important educational resources, individuals from these groups may be likely to have achieved lower levels of ability as reflected on important educational indicators

(e.g., SAT scores, etc.). This can be true even if these individuals have the same level of capacity to learn these abilities in the future as do individuals from other socioeconomic groups. Thus, “abilities measured at a single point in time, no matter how reliably measured or sophisticatedly modeled, are not synonymous with potential,” even though they are often treated as such (McNeish & Dumas, 2017, p. 61). In addition to the potentially harmful consequences of these biases, this also means that the results are simply less accurate—the results of the assessments are not accurately reflecting the constructs and goals for which they are being used, yielding a threat to utility and to construct validity.

This issue is of particular importance in educational testing, but it is not unique to that field. In the field of medical diagnostics, for example, Raket (2020) explicitly argued for the use of mixed effects growth/decay modeling in Alzheimer's diagnosis. He argued that an individual’s cognitive ability (which is used to diagnose Alzheimer’s Disease) at any given point in time is partially a function of that person’s maximum cognitive ability: patients with greater cognitive abilities before the onset of the disease are also likely to score better on cognitive tests in the early stages of having the disease than are those patients with lower starting cognitive abilities. Therefore, testing at a single time-point may over-diagnose those patients who had lower starting cognitive capabilities and under-diagnose those who had higher starting cognitive capabilities, indicating a significant inadequacy with single-timepoint testing (Raket, 2020, p. 1).

In the field of education, several attempts have been made to rectify this by moving beyond single-time-point assessments. In the decades after World War II,

dynamic assessment was developed as a way to assess the child survivors of concentration camps (Feuerstein et al., 1979; 2015). Students were tested multiple times with targeted learning opportunities in between assessments. The improvement over time was plotted to identify a *capacity* value representing the expected future ability once the construct of interest was fully developed. The use of dynamic assessment continues today in certain contexts such as second language acquisition (Lantolf & Poehner, 2011), students with intellectual disabilities (McLaughlin & Cascella, 2008), intellectually gifted students (Kirschenbaum, 1998), and others.

While dynamic assessment directly addresses the problem inherent in single-administration tests, it is also difficult to implement on a large scale due to the high level of resource commitment required. The targeted learning opportunities in between assessments call for one-on-one instruction, requiring significant amounts of time and funding. Not only is this difficult to scale to large populations, it also makes instructional standardization difficult. Furthermore, dynamic assessment has typically used descriptive plots of individual student growth rather than any formalized statistical model or growth curve (McNeish & Dumas, 2019; McNeish et al., 2020), which also makes scaling up to large populations difficult and resource-intensive.

More recently, DMM models, building on the conceptual framework of dynamic assessment, were developed as a solution to the problem. Like dynamic assessment, they use longitudinal data to directly estimate an individual's capacity. Unlike dynamic assessment, though, they rely on statistical models rather than on descriptive plots. DMM models therefore build on dynamic assessment by incorporating dynamic assessment's

conceptualization of student growth capacity within a statistical model; DMM models seek to estimate student capacity as in dynamic assessment, but they aim to do so with large longitudinal datasets and without the need for personalized one-on-one interventions or instruction (Dumas et al., 2020, p. 286).

With multiple datapoints (scores) over time, the model estimates a non-linear growth trajectory (the shape of which is typically specified by the user). The capacity is then typically conceptualized as the upper asymptote to this growth trajectory. By adding a random effect on this asymptote, each individual can have his or her own asymptote and therefore his or her own capacity estimate. Thus, DMM models are effectively nonlinear mixed effects models with a random effect on an upper asymptote parameter. This individual-level capacity estimate is equivalent to the capacity estimate from dynamic assessment, but it is estimated directly by the model, allowing thousands of students to each receive such an estimate without clinicians needing to subjectively analyze each individual's growth trajectory. DMM models can therefore more readily be scaled and used with large groups of students, with a relatively smaller resource commitment, than dynamic assessment.

Even though DMM models are specified in a way that mimics a typical nonlinear growth model, their application and interpretation are much more like an item-response theory (IRT) model: each individual has an unobserved capacity for learning the particular ability of interest. The random effect on the upper asymptote estimates this latent capacity for each individual. DMM models can also have various other individual-level random effects, depending on the shape. These can include growth rate parameters,

midpoint parameters, and inflection point parameters. However, since the goal is typically to estimate a student's capacity, the random effect on the upper asymptote (or some other parameter representing the function's maximum value) is the most substantively important of these estimates in the DMM context. Due to their original intent of providing a capacity estimate for students' growth on a given educational ability, DMM models are typically implemented and interpreted in this educational context. However, such models could theoretically be applied to any situation in which a capacity estimate (an individual-level upper asymptote) on a nonlinear growth trajectory would be of substantive interest (e.g., estimating an individual's maximum skill in a particular sport, estimating a tree's maximum height at maturity, estimating a puppy's eventual weight as an adult dog, or estimating a student pilot's capacity for flying ability).

Importantly, in the educational context, these models have been shown to improve consequential validity in comparison to single-administration scores. Specifically, DMM capacity scores have been more weakly correlated with demographic variables than single-administration scores are. Dumas and McNeish (2017) demonstrated this in the context of mathematics ability, while Dumas and McNeish (2018) demonstrated a similar result with reading ability: socioeconomic status (SES) predicted about 20% of the variance in reading scores at individual time points ($R^2 \approx 0.20$), while SES predicted only 3.4% of the variance in capacity estimates ($R^2 \approx 0.034$) produced by a DMM model. Thus, the DMM capacity estimates are potentially less susceptible to the biases that exist in high-stakes single-administration tests.

Dumas and McNeish (2017) also showed that the capacity estimates from DMM models are reliable over time. When they removed the last time point and re-ran the analysis, the restricted-model capacity score estimates correlated well above 0.90 with the full-model capacity score estimates. This evidence of reliability is important, as it demonstrates that the capacity score estimates are not heavily dependent on which time points are available. Thus, the capacity score estimates are not just computationally expensive noise; they are instead likely measuring an extant construct of capacity. Further evidence of this comes from McNeish et al. (2020), who showed that adult scores on a verbal ability test were much more accurately predicted by the capacity scores from a DMM model of adolescent scores than by a longitudinal IRT model (R^2 of 0.43 versus 0.16).

Since their relatively recent inception, DMM models have been applied in several educational contexts. These include verbal ability (McNeish et al., 2020), mathematics assessment (Dumas & McNeish, 2017; Dumas et al., 2020;Dumas, McNeish, Sarama, et al., 2019; Dong et al., 2022), reading ability (Dumas & McNeish, 2018), summer learning loss (McNeish & Dumas, 2021), and medical education (Dumas, McNeish, Schreiber-Gregory, et al., 2019). Further research has created additional quantities of interest that can be calculated from a DMM model. Dong et al. (2022) developed an individual-level trajectory deviance index (TDI) that can be used to assess how well a DMM model fits for each individual in the dataset. McNeish and Dumas (2018) developed a conditional reliability measure for DMM, allowing researchers and practitioners to assess the consistency of estimates from DMM models.

DMM growth curves have typically followed two broad classes of growth trajectories. The J-shaped growth trajectories exhibit growth in which the estimated scores are always increasing smoothly at a decreasing rate as a function of time. Thus, for the J-shaped curves, the low end of the curve is conceptualized as an intercept (predicted score at time zero). The S-shaped growth trajectories, by contrast, have both lower and upper asymptotes and exhibit an inflection point somewhere in between the starting score and the upper asymptote. The low end of the curve is conceptualized as a lower asymptote rather than as an intercept for these S-shaped curves.

McNeish et al. (2020) identified six different curve trajectories that can be parameterized to have an upper asymptote. Four of these (Michaelis-Menten, Exponential, Gompertz, and Morgan-Mercer-Flodin) are J-shaped curves with an intercept parameter, while the other two (Logistic and Weibull) are S-shaped curves with lower asymptotes. Even within a given shape, there are multiple different parameterizations that could be used (see Preacher & Hancock, 2015; Tjørve & Tjørve, 2017).

For the purposes of illustration, four of these DMM growth curve shapes are shown in Table 1 below: two J-shapes and two S-shapes. The two included J-shaped curves are Michaelis-Menten and Exponential, while the two included S-shaped curves are Logistic and Weibull. The exact parameterizations and parameter definitions for each of these shapes can be found in Table 1. Note that, due to the many parameterization options for each trajectory, some parameterizations listed here may differ slightly from

those used in previous DMM implementations and from those in other nonlinear growth models more generally.

Table 1: Example of DMM Growth Trajectories

Curve	Parameterization	Parameter Definitions
Michaelis-Menten	$\beta_0 + \frac{(\beta_c - \beta_0) * t}{\beta_m + t}$	β_0: Intercept β_c: Capacity β_m: Midpoint t: Time
Exponential	$\beta_0 + (\beta_c - \beta_0)(1 - e^{\beta_r * t})$	β_0: Intercept β_c: Capacity β_r: Growth rate t: Time
Logistic	$\beta_L + \frac{\beta_c - \beta_L}{1 + e^{\beta_r * (t - \beta_m)}}$	β_L: Lower Asymptote β_c: Capacity β_r: Slope at midpoint β_m: Midpoint t: Time
Weibull	$\beta_c + (\beta_L - \beta_c) * \left(1 - e^{-e^{\beta_r * (\ln(t) - \ln(\beta_i))}}\right)$	β_L: Lower Asymptote β_c: Capacity β_r: Growth rate β_i: Inflection Point t: Time

Each of these shapes contains either a lower asymptote (β_L) or an intercept (β_0). Even though they technically have different substantive interpretations, in many settings they are effectively equivalent, since they both represent the lowest value of the vertically scaled outcome value that we would expect to see. The Michaelis-Menten and Logistic trajectories both have midpoint parameters (β_m), which represent the point on the time scale where the outcome is halfway between the upper asymptote and either the lower asymptote (Logistic) or intercept (Michaelis-Menten). All of the trajectories except for Michaelis-Menten have a growth rate parameter (β_r); in the Logistic curve, this takes on particular meaning as the slope at the midpoint. The Weibull trajectory is the only one with an inflection point parameter—although there is an inflection point in the Logistic trajectory, it is not directly estimated as a parameter. Implied by the nature of DMM, all four shapes have an upper asymptote representing capacity.

The parameters listed in Table 1 all represent fixed effects. They represent the estimated average for that parameter across the participants. For example, if $\beta_L = 3$, then the average estimated lower asymptote for the participants is three. Regardless of the trajectory, each of the parameters in that curve can be given a random effect. This allows each individual to have his or her own estimated value for that parameter. Since Time is not a parameter to be estimated, it cannot have a random effect.

It is up to the researcher/practitioner to decide (whether *a priori* or based on model fit) which parameters should be given random effects. However, the entire concept of DMM relies on each individual being given his or her own capacity score. It is therefore highly recommended that the β_c parameter receive a random effect, while the

decision to utilize a random effect on the other parameters may depend on theory, model fit, and/or model convergence.

Similarly, the choice of which DMM trajectory to use may be inductive or deductive. Deductively, the researcher/practitioner may have *a priori* expectations about the shape of the growth curve, particularly about whether to expect a J-shaped curve, an S-shaped curve, or some other trajectory. A specific growth trajectory may also be preferred for various other reasons (interpretability, presence of a particular parameter, etc.) as well. This approach was taken by McNeish and Dumas (2019), who decided *a priori* to use a Michaelis-Menten trajectory due to its ease of interpretation. Conversely, because DMM users are primarily interested in the capacity estimates on the upper asymptote, the researcher/practitioner may not have much preference or expectation concerning which trajectory to use. In this case, the researcher/practitioner may try each of the trajectories (possibly along with different combinations of random effects) to see which ones are able to achieve convergence and which one has the best fit. McNeish et al. (2020) employed this technique, using BIC to compare the fit of Michaelis-Menten, Exponential, Logistic, and Weibull models.

To date, all of these DMM models and associated quantities of interest have been estimated using *SAS*® `PROC NLMIXED` (*SAS Software*, n.d.).[8] However, nonlinear mixed effects models can be conducted in *R* (Team, 2022; R Core Team, 2022) as well by using the `nlme` package (J. Pinheiro et al., 2022; J. C. Pinheiro & Bates, 2000) and/or

[8] *SAS* and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

its relative, the `lme4` package (Bates et al., 2015), each of which has thousands of citations. The main benefit of this is that *R* is free and open-source, providing access to those who may not have the resources or institutional backing to access *SAS* software.

In addition to the monotonically increasing J-shaped and S-shaped growth curves that have traditionally been implemented in DMM models, other growth curve shapes are possible. For example, "the linear growth model is the most commonly fit growth model in the social sciences" (Grimm et al., 2016, p. 201). However, a linear growth model cannot have a meaningful measure of capacity, as each person's ability estimate will simply continue to increase at a constant rate as a function of time, thereby leaving each individual with an infinite capacity estimate. Nonetheless, some other shapes can be reinterpreted as DMM models. For example, quadratic growth does not have an upper asymptote, but it does have a maximum or minimum value at the vertex of the quadratic function. With a random effect on the vertex, each individual could have his or her own value for the vertex (both his or her own point in time at which the vertex occurs and his or her own maximum or minimum value). This could readily be interpreted as a capacity estimate. Along these lines, McNeish et al. (2022) discuss the relationship between quadratic growth functions and DMM models; they end up building a piecewise growth function that combines a quadratic growth function on the left side of the maximum knotted with a horizontal linear function beginning at the maximum. Although the authors discuss this model as borrowing aspects of dynamic measurement (as opposed to being a DMM model itself), one could reasonably view this as a new variant of DMM. The maximum ability level is directly estimated by the model and receives a random

effect, allowing it to be interpreted as an individual-level capacity estimate. A pure quadratic model would have this same characteristic: with the right parameterization, the maximum ability level can be directly estimated and can receive a random effect, allowing each individual to have his/her own capacity estimate.

Grimm et al. (2016) used the standard form equation to represent quadratic growth: $Y_{ti} = b_{1i} + b_{2i} * t + b_{3i} * t^2 + u_{ti}$ (Grimm et al., 2016, p. 203). However, this version of the equation does not directly show the vertex as a parameter, so it is not possible to add a random effect on it. Instead, other parameterizations would be more helpful from a DMM perspective. McNeish et al. (2022) discuss some options for parameterizing quadratic growth functions. Conceptualizing the quadratic growth equation in vertex form would be one useful option from a DMM perspective: $Y_{ti} = a_i(t - h_i)^2 + k_i$.[9] Now the h and k parameters would represent the time at which the function reaches its vertex (presumably a maximum) and the value of that maximum, respectively. In the context of DMM, then, they would represent the time/age at which an individual reaches his or her capacity and the value of that capacity, respectively.

[9] Note that this is similar to, but slightly different from, the parameterization selected by McNeish et al. (2022).

Chapter Three: Methods

To measure each golfer's ability in a given year, I used a linear mixed effects model with a player's score in a given round as the outcome variable. I extracted the estimates for each player's ability in a particular year from this model. Thus, almost all players received multiple ability estimates over time. I then used the multiple longitudinal player-level estimates to estimate multiple longitudinal growth trajectories that allow for a capacity parameter to be directly estimated, selected the best-fitting of these (the quadratic model), and created a Dynamic Measurement Model using the selected growth trajectory. Finally, I used this DMM model to create actual capacity estimates for specific professional golfers and validated the results.

Data

All data that were used are publicly available. I collected the names and scores of all participants in official events on the PGA Tour from 2007 to 2023. I also collected the names and scores of all participants in official events on professional golf tours that are frequently crossed with the PGA Tour (i.e. there is significant overlap between these tours, with players frequently moving between them and/or playing events on both tours): the Korn Ferry Tour (previously called the Nike Tour, the Buy.com Tour, the Nationwide Tour, and the Web.com Tour), PGA Tour Champions (previously called the Senior PGA Tour and the Champions Tour), PGA European Tour (also called the DP World Tour and

the European Tour), and LIV Golf (founded in 2021 as a direct competitor to the PGA Tour). For each player with at least four rounds of golf in the dataset, I also collected (or attempted to collect) his date of birth in order to calculate age. Additionally, for each round of golf played, I collected the name of the golf course. Thus, for any given player-round observation, the only variables collected are the player's name, the name of the golf course, the player's date of birth (to calculate age), the date/round number in which the round takes place, and the player's score in that round. All of these data are publicly available online and are not private; no interaction with participants was required.

The final dataset resulted in 866,539 rounds of golf played by 8,397 professional golfers across 528 golf courses. This yielded a mean number of rounds per player of just over 103 rounds. This also yielded 33,236 player-year combinations. However, this number decreased slightly after dropping players with fewer than four rounds of golf in the dataset and dropping those whose dates of birth could not be ascertained. The result was 29,423 player-year combinations across 5,225 players, yielding an average of about 5.6 years in the dataset per golfer.

Summary of Variables

For the primary analyses, I collected only six variables: the name of the player, the year in which the round occurred, the course on which the round occurred, the round number within the tournament (first round, second round, etc.), the score that the player shot in the round, and the player's date of birth. Two variables were directly calculated using these six variables: the number of shots behind the lead a player was after the previous round and the player's age. Two other variables were collected for the purposes

of validation: the player's OWGR ranking at the end of the previous year and the player's (total) Strokes Gained per round at the end of the previous year. Two variables were estimated by the primary analyses in this study: player-year ability estimates from the linear mixed effects model and player capacity estimates from the DMM model. Finally, several variables were estimated by ancillary models used for validation: ability estimates from the SBSE model, pooled age-invariant player ability level estimates, and player capacity estimates from restricted early-career models. A summary of all of the variables collected, calculated, or estimated for use in primary analyses can be found in Table 2.

Linear Mixed Effects Model to Measure Golfing Ability

Primary Analysis

The linear mixed effects model to measure each player's golfing ability in a particular year took the form of a crossed linear mixed effects model. The unit of analysis was a round of golf, with the outcome variable being the score that the player achieved in that round. In its most basic form, such a mixed effects model could include merely a fixed effect intercept term with random effects for the player and the course. However, such a model would assume that each course does not vary in difficulty from round to round. Thus, it is likely necessary to include an additional random effect for the course-day or course-round to account for variations in course difficulty from day to day arising from course setup, weather, etc. Such a model would be similar to the SBSE ("Score-based Skill Estimate") Model presented by Broadie and Rendleman Jr, (2013).[10]

[10] Although their notation is consistent with a mixed effects model, Broadie & Rendleman Jr (2013) actually appear to have used dummy variables instead. The logic of the model is the same, however. I used the true mixed effects version.

Table 2: Variables Collected or Calculated for Primary Analyses

Variable Name	Description	Level of Measurement	Data Source	Primary Use
Player Name	The player playing the given round of golf	Round	Public data (internet)	Grouping variable to create player-level ability estimates and player-level capacity estimates
Year	The year in which the round was played	Round	Public data (internet)	Grouping variable to create player-year ability estimates and course-year difficulty estimates
Course	The course on which the round was played	Round	Public data (internet)	Grouping variable to create course-level and course-round difficulty estimates
Round Number	Which round of the tournament is being played	Round	Public data (internet)	1. Grouping variable to estimate course-round difficulty estimates 2. Fixed effect for interaction terms with shots behind lead
Round Score	The score achieved by the player in the given round	Round	Public data (internet)	Outcome variable in linear mixed effects model
Date of Birth	The player's date of birth	Player	Public data (internet)	Used to calculate player's age
Shots behind Lead	The number of shots behind the lead a player is at the conclusion of the previous round	Round	Calculated based on previous round score and minimum score within tournament-year	Predictor variable in linear mixed effects model
Age	The player's age	Player-year	Calculated based on Year and player's Date of Birth	Used as input variable in DMM model
Player-Year Ability	Estimated ability level for the player in the given year	Player-year	Estimated by the linear mixed effects model	Used as outcome variable for DMM model

However, I also added fixed effects for how far behind the leader a player was before starting a round, the squared version of this variable, and the interaction of both of these with the round number. This allows the model to account for the possibility that players might outperform or underperform their ability levels due to stress and nerves (if they are near the leader) or apathy (if they are far away from the leader) as well as the possibility that these effects will vary depending on how early or late it is in the tournament. Even if these variables were to have been statistically nonsignificant predictors, then there would have been no harm in including them: their inclusion would not have biased the estimates of the random effects. However, their inclusion prevented the possible bias that may have resulted from these effects being misattributed to player ability.

I also included three additional random effects. First, I added a player-year random effect, which allows player ability to vary from year to year instead of running an entirely separate model for each year or for each two-year period. Second, I included a course random effect to supplement the course-round random effect. This provided an overall measure of each course's mean difficulty level, allowing the course-round random effect to represent deviations from this overall mean difficulty due to weather conditions or course conditions. Third, because these deviations in course difficulty may be more similar to each other within a given year than they are across years, I also added a course-year random effect. The resulting linear mixed effects model that I used took the following form:

$$S_{icdy} = \beta_0 + \beta_1 L_{icdy} + \beta_2 L_{icdy}^2 + \beta_3 R_{dy} + \beta_4 \left(L_{icdy} * R_{dy}\right) + B_5 \left(L_{ijdy}^2 * R_{dy}\right) + \kappa_c + \delta_{cd} + \eta_{cy} + \theta_{iy} + \upsilon_i + \varepsilon_{icdy}$$

In this equation, subscript i represents the individual golfer, subscript c represents the course, subscript d represents the day or round, and subscript y represents the year. Thus, S_{icdy} as the outcome variable means the score of player i on course c and day/round d of year y. Thus, the random effect κ_c represents a golf course's mean difficulty, υ_i represents a player's mean ability level across years, and θ_{iy} represents a player's deviation from that mean ability level in a particular year. Given this, a player's ability level in a particular year can be represented by $\upsilon_i + \theta_{iy}$, where lower values represent higher ability (because lower scores are better in golf).

The beta coefficients represent the fixed effects. β_0 represents the mean score of an average player who is 0 shots behind the lead in (the theoretical) round 0. β_1 and β_2 represent the increase or decrease in the score from the mean per shot that a player is behind the leader and per squared shot that a player is behind the leader, respectively. β_3 represents the increase or decrease in score as the tournament progresses—this is not of much use on its own, but it is included because it is later used in interaction terms. β_4 and β_5 represent the extent to which the effect of $\beta_1$and β_2 vary depending on the round number (how late or early it is in the tournament).

From this model, I extracted the estimates of each player's ability in each year ($\upsilon_i + \theta_{iy}$). Since most people intuitively think of higher values as being qualitatively

better in terms of ability, I multiplied these values by negative one so that $-(\upsilon_i + \theta_{iy})$ now represents an ability measurement where higher abilities are represented by larger (positive) numbers. Unlike those from many measurement models, these ability values actually are on a substantively meaningful scale: the number of strokes above or below average a player is. Thus, an ability level of -.56 would mean that the player's ability level is about half a shot per round worse than the average professional golfer on a major tour.

Validity Analysis

I validated the player-year ability level estimates in two ways. First, I calculated the correlations between the new scores and three existing unidimensional measures of golfing ability—the Official World Golf Rankings (OWGR), Total Strokes Gained (per round), and estimates from the SBSE model created by Broadie and Rendleman (2013)—as well as with pooled, age-invariant player ability estimates created by estimating a separate model with only player-level random effects (no player-year random effects). The OWGR began in the 1980's and encompasses all professional golf tours, and SBSE and the pooled, age-invariant estimates can be estimated for any time period, allowing both of these measures to be calculated for the entire sample. However, the PGA Tour is the only tour to track Strokes Gained, meaning that this correlation was performed on a restricted sample limited to golfers playing enough rounds of golf on the PGA Tour in a given year for the PGA Tour to provide Strokes Gained data. The expectation was that the newly estimated player-year ability scores would be positively, but not perfectly,

correlated with the existing unidimensional measures (hypothesized magnitudes in the 0.6 to 0.8 range). Such correlations would provide evidence of convergent validity.

The new score estimates and the SBSE estimates are naturally player-year level estimates, and total Strokes Gained is provided at this level as well (in addition to being provided at more granular player-round or player-tournament levels). However, OWGR fluctuates weekly based on the most recent tournament results. As such, it requires a bit of adjustment to be comparable to the other measures. To deal with this, I treated the last OWGR of the season as the official player-year level measure of ability for the given year.

Second, I compared the predictive ability of the new estimates to the predictive abilities of the existing unidimensional measures of golfing ability. This took the form of multiple linear mixed effects regressions. These regressions still had random effects for course, course-round, etc, as did the primary analysis in this stage, but these validation analyses used the respective measures of ability from the previous year rather than any of the player-level random effects. For example, the five ability measures (including the new estimates from the primary analysis) from 2010 were used to predict player scores in 2011. The predictive ability of each measure was assessed using mean square error—lower mean square error of prediction values indicate better predictive ability. For the new score estimates to be accurate, they should perform at least as well as the other existing measures of ability at predicting future performance. This would provide evidence of predictive validity as well as of incremental validity. Dong and Dumas (2024) discuss this idea of incremental validity as a critical component of DMM validity.

Dynamic Measurement Model(s) To Estimate Players' Capacities

Primary Analysis

In the second stage of the analysis, I took the player-year ability level estimates from the previous stage and used them in DMM models that are able to estimate each player's capacity—the highest possible ability level that the player is predicted to achieve. For some players, this value may be in the past. From the standard DMM models, I estimated two J-shaped curves (Michaelis-Menten and Exponential) and two S-shaped curves (Logistic and Weibull).

In addition to the more common DMM shapes that monotonically increase or decrease, I also used quadratic growth. Quadratic growth allows for ability to first increase and then decrease, which may be useful for modeling golfing ability, as we may expect individuals to reach their capacities at a relatively young age. Even though it has rarely been previously used in DMM models, quadratic growth is a commonly used longitudinal growth curve shape (Grimm et al., 2016). Grimm et al. (2016) use the standard form equation to represent quadratic growth: $Y_{ti} = b_{1i} + b_{2i} * t + b_{3i} * t^2 + u_{ti}$ (Grimm et al., 2016, p. 203). However, I used a different quadratic parameterization that shows the coordinates of the vertex directly as parameters: $Y_{ti} = a_i(t - h_i)^2 + k_i$. This is more helpful from a DMM perspective, as the value at the vertex represents the individual-level maximum ability (i.e. capacity). Converting this even further so that the parameterization looks more like the more traditional DMM equations, we could write this as $Y = \beta_r(t - \beta_a)^2 + \beta_c$, where β_r is a growth rate parameter, β_a is the time at the vertex (i.e. the age at which a player reaches his capacity), and β_c is the player's capacity

estimate. This is similar, but not identical, to the quadratic parameterization that McNeish et al. (2022) used as a portion of a piecewise nonlinear growth trajectory. I estimated these DMM models in *R*, and I provide the *R* code for doing so in Chapter Four,[11] providing a tutorial for future DMM users. The parameterization of each of the growth trajectories to be estimated can be found in Table 3.

Table 3: Growth Trajectories to be Estimated

Curve	Parameterization	Parameter Definitions
Michaelis-Menten	$\beta_0 + \frac{(\beta_c - \beta_0) * t}{\beta_m + t}$	β_0: Intercept β_c: Capacity β_m: Midpoint t: Time
Exponential	$\beta_0 + (\beta_c - \beta_0)(1 - e^{\beta_r * t})$	β_0: Intercept β_c: Capacity β_r: Growth rate t: Time
Logistic	$\beta_L + \frac{\beta_c - \beta_L}{1 + e^{\beta_r * (t - \beta_m)}}$	β_L: Lower Asymptote β_c: Capacity β_r: Slope at midpoint β_m: Midpoint t: Time
Weibull	$\beta_c + (\beta_L - \beta_c) * \left(1 - e^{-e^{\beta_r * (\ln(t) - \ln(\beta_i))}}\right)$	β_L: Lower Asymptote β_c: Capacity β_r: Growth rate β_i: Inflection Point t: Time
Quadratic	$\beta_r(t - \beta_a)^2 + \beta_c$	β_r: Growth rate β_a: Time at capacity β_c: Capacity t: Time

[11] Demonstrating *R* code in the body of the text for future reference by readers is similar to the approach taken by Carsey & Harden (2013).

Before estimating the true DMM model, I began by estimating the marginal (i.e. fixed effects only) model for each growth trajectory. The best-fitting of these marginal models (the quadratic model) was selected for further analysis with random effects. This inductive approach is similar to that of McNeish et al. (2022). The best-fitting model was selected using two criteria: Bayesian Information Criterion (BIC) and Mean Square Error (MSE). Lower values of each of these represent better fit. Both of these criteria have been used previously to select DMM trajectories: McNeish et al. (2022) used MSE to select a growth trajectory, while McNeish et al. (2020) used BIC to select a growth trajectory.

Once the best-fitting growth trajectory had been selected, I converted it into a true DMM model by adding random effects to each parameter. I then iteratively dropped random effects (except for the capacity parameter), noting which models achieve convergence and which have the best fit. The best fitting model that achieved convergence was selected. Consistent with the conceptual underpinnings of DMM, I always retained a random effect on the capacity parameter. The best-fitting DMM model then became the final model used to estimate player-level capacity scores.

Validity Analysis

To validate the DMM model as a measurement model, I followed many of the suggestions of Dong and Dumas (2024). In particular, they argued for the importance of incremental validity in the context of Dynamic Measurement. From this perspective, validity may be thought of as improved predictive ability beyond what has been provided by existing measures or models. Thus, the primary goal of this DMM validation was to assess whether the DMM model added value in the form of improved predictive ability

compared to existing measures of golfing ability/capacity. I proceeded with three distinct validation steps: assessing the relationship between the full-model DMM capacity estimates and those of restricted models, comparing the fit of the nonlinear DMM growth trajectory to that of a more naïve linear growth trajectory, and testing whether DMM capacity scores can outperform three baseline models in forecasting players' future capacities.

Correlation with Restricted Models

With 17 years of data in the sample in this project, the sample captured the entire careers of some players, while it likely only included the beginnings or the ends of the careers for other players. From a practical perspective, however, many future DMM users will only have data from the early datapoints in an individual's growth; they will want to predict future capacity using these early values. Thus, a valid DMM model should be able to forecast future capacity accurately without access to data at those future time points. To assess the extent to which this was occurring, I ran the same model on a restricted-sample (early career) dataset, and I calculated the correlation between the restricted-sample capacity estimates and the full-sample capacity estimates. A strong positive correlation provided evidence for the utility and validity of the capacity score estimates, demonstrating that the estimates are not solely dependent on which datapoints are available.

Similarly, I also performed a modified version of 10-fold cross-validation.[12] This method of validation splits the dataset into 10 subsets. In a typical k-fold cross-validation context, one of the 10 datasets is withheld, the model is run on the other nine datasets, and then it is tested on the withheld dataset. This is then repeated for each of the 10 subsets. However, with a DMM model, there is no good way to 'test' the capacity estimates for the withheld dataset, especially if any given player's rounds are not distributed across the 10 subsets. This is a difficulty with mixed-effects models in general. Because of this, I ran a modified version of the cross-validation. For each of the subsets of data, I ran the same model as the full-model, extracted capacity estimates, and calculated the correlation between these new capacity estimates and the full-model capacity estimates (dropping any player that did not appear in both datasets). As with the correlation between the full-model capacity scores and the age-based restricted model, a strong correlation between the full-model capacity scores estimates and the k-fold capacity score estimates provided evidence of predictive validity.

Comparison to Linear Growth Trajectory

A linear growth model cannot produce individual-level capacity scores (as they would all be infinite), but it is still possible that such a growth model would provide a good description of players' growth trajectories over the course of their careers. Since such models are significantly less complex to run, a nonlinear DMM model would ideally provide noticeable improvements over a linear model in order to make it worth the extra

[12] For k-fold cross-validation, k=5 or k=10 is recommended by Kuhn and Johnson (2013) and James et al. (2021). Carsey and Harden (2013, pp. 259–262) provide example *R* code for performing cross-validation.

human effort and extra computational effort. To test this, I applied a linear growth model to the longitudinal ability scores, including random effects for the slope and the intercept. I then compared the model fit of the selected DMM model to that of the linear growth model. In the end, the DMM model and the linear model showed roughly equal fit, though the linear model did not converge in *R*. Although this was weaker-than-expected support for the DMM model over the linear model, the DMM model was still somewhat preferable to the linear growth model.

Comparison to Single-Timpeoint Estimates of Capacity

One benefit of DMM capacity estimates is that they reduce the reliance on single-timepoint scores as implicit indicators of future success. This has been shown in educational settings, and I tested whether it holds true in the measurement of golfing ability. I tested the capability for the capacity scores produced by the DMM model to forecast future ability in comparison to single-timepoint early-career ability estimates.

The ages/years to be used in the restricted model were selected based on the results from the DMM model. Because players reached their peak around age 32 (on average) according to the DMM model, I used ages 30 and lower for the validation. I used each age as a separate single-timepoint predictor: the ability for age 18 scores to predict the player's maximum player-year ability, the ability for age 25 scores to predict the player's maximum player-year ability, etc.

Each of these was treated as an independent variable in a basic linear regression with the player's highest player-year ability score estimate from the linear mixed effects models used as the dependent variable. I ran similar basic linear regressions with each

possible pre-maximum age as the independent variable and compared each of them to a model in which the DMM capacity scores were used to predict the player's maximum player-year ability level from stage one. To show incremental improvement in validity over these single-timepoint estimates of capacity, the DMM capacity score needed to demonstrate a higher predictive ability (higher R^2) than the single-timepoint estimates, especially when the capacity score estimate was generated on an age-restricted dataset.

Similarly, I followed the same procedures using single-year OWGR and single-year Total Stroked Gained values from early in players' careers as single-timepoint measures of ability/capacity. Each of these was used as the predictor variable in basic linear regressions, with the player's highest player-year ability estimate as the dependent variable. The DMM capacity estimate needed to produce a higher R^2 than these single-timepoint measures of ability/capacity in order to provide validation.

Tutorial for Conducting DMM in *R*

While estimating the DMM portions of the analysis, I embedded *R* code and output into the body of the text to demonstrate that the DMM models can be run in in *R* and to provide a tutorial for how to conduct DMM models in *R* for the benefit of future DMM users. DMM models to date have been run almost exclusively in *SAS*, making the ability to run them in *R* a significant step forward. I have also provided commentary on the code and syntax being used, the options being selected, and the output being produced. The embedding of code and output will be enabled by *R Markdown*, which allows for the creation of PDF, HTML, and MS Word documents with *R* code and output embedded (Baumer & Udwin, 2015).

Although it would be possible to embed *R* code for all primary and validation analyses as well, many of these procedures (e.g. linear regression) are well-documented and frequently used in *R*. Thus, there is no need for a new tutorial on these methods, and my inclusion of such syntax would not provide much added benefit. For the sake of parsimony, then, I have only provided coding tutorials for the DMM portions of the analyses.

Chapter Four: Results

The results are provided in two broad sections, each representing one of the two stages of the analysis. The first section, providing the results of the stage one analysis, shows the results of the linear mixed effects model and the validation of this model. This stage used professional golfers' scores to estimate their ability levels in specific years and then validated these results. The results were favorable: the player-year estimates were correlated with existing measures of golfing ability, but they provide better predictions (lower MSE) of the observed scores.

The second section provides the results of the DMM model and the rest of stage two. In this stage, multiple potential longitudinal growth models were fit as marginal models (fixed effects only), and the best-fitting of these was used as the trajectory for the DMM model, which adds random effects to the marginal model. Each potential trajectory had a parameter that can be conceptualized as a capacity score. By adding a random effect to it, each individual golfer can receive his own capacity score representing his maximum forecasted ability level throughout his career. These results were then validated.

The results indicate that a quadratic model is the best fitting growth trajectory according to both BIC and MSE, that the capacity score estimates from this model are reliable (not dependent on particular timepoints), that the quadratic DMM model provides

(slightly) better fit than a baseline linear mixed effects growth model, and that the capacity scores from the DMM model are better able to predict player-year maximum scores from stage one than are single-timepoint early-career estimates of ability.

Stage One: Linear Mixed Effects Model

Primary Analysis

The primary goal of stage one was to generate player-year ability estimates that could then be used in stage two. This was accomplished via the estimation of a crossed linear mixed effects model. Nonetheless, the model itself produced results that may provide some insight into the factors that affect a player's score in a given round of golf. Results from this model can be found in Table 4.

All of the fixed effects achieved statistical significance, indicating that situational factors, such as the number of shots behind the leader a player is at a given point in the tournament, can affect a player's performance. However, because of the interaction effects, the magnitude of these effects is easier to see in a graph or table. Table 5 shows the magnitude of these effects by showing the expected score for an average professional golfer on a major tour on a course of average difficulty in different situations.

Table 4: Stage One Linear Mixed Effects Model

Fixed Effects	Estimate	SE	t
(Intercept)	73.90	0.07159	1032.318
Shots Behind Leader	0.04378	0.003454	12.675
(Shots Behind Leader)^2	-0.0002316	0.00002909	-7.959
Round #	-0.07648	0.01017	-7.518
Shots Behind * Round #	-0.01027	0.001116	-9.202
(Shots Behind)^2 * Round #	0.00005883	0.000008134	7.233
Random Effects	Variance	SD	
Player	5.6093	2.3684	
Player-Year	0.4353	0.6597	
Course	1.6843	1.2978	
Course-Year	0.4829	0.6949	
Course-Year-Round	0.6560	0.8099	
Residual	8.0022	2.8288	

While these effects are statistically significant, Table 5 shows that their magnitude is relatively small. Even the smallest expected score (a player leading the tournament in round 4) and the greatest expected score (a player 25 shots behind the lead in round 2) are less than a full shot apart. Thus, statistical significance here is likely due in part to the large sample size and does not necessarily indicate substantive significance. However, the fixed effects were included primarily as control variables to eliminate bias, so their substantive interpretation is less important.

Table 5: Expected Scores in Various Situations

	Round 1	Round 2	Round 3	Round 4
Leader	73.82	73.75	73.67	73.59
1 Shot Behind		73.77	73.68	73.60
2 Shots Behind		73.79	73.70	73.60
3 Shots Behind		73.82	73.71	73.60
4 Shots Behind		73.84	73.72	73.60
5 Shots Behind		73.86	73.73	73.61
7 Shots Behind		73.90	73.76	73.61
10 Shots Behind		73.97	73.79	73.62
15 Shots Behind		74.07	73.85	73.64
20 Shots Behind		74.17	73.91	73.65
25 Shots Behind		74.26	73.96	73.66

More importantly, the random effects show that the player's ability level and random variation (luck) are by far the most important factors determining the score in a given round of golf. The random effect for player accounts for a variance of 5.6093, much greater than all of the other random effects other than the residual. The standard deviation (the square root of the variance) is more interpretable: a player who is one standard deviation above average would be expected to score 2.3684 better, on average, than an average professional golfer on a major tour. The residual standard deviation is even larger: on average, a player would be expected to deviate from his "true" expected

score in a given round by over 2.8 shots just due to random variation. This is consistent with prior studies showing residual standard deviations between 2.69 and 3.12.

These residuals appear to be approximately normally distributed (see Figure 1), and Figure 2 shows that there does not appear to be any pattern in the residuals across the years in the dataset (2007-2023). Both of these provide confidence that the model assumptions and conditions for inference have been met and lend credibility to the predictions that the model has generated. I therefore proceeded with generating player-year ability estimates from the model and performing more formal validation steps.

Figure 3 shows the distribution of player-year ability estimates (after multiplying by negative one so that positive values indicate greater ability) from the linear mixed effects model. Interestingly, there is a clear skew to the left in the distribution, indicating that a small proportion of player-year observations in the dataset are significantly 'worse' (lower ability) than the rest of them. These may be players who received sponsors invitations into tournaments and did not qualify through performance. Conversely, the best player-year observations are relatively closer to the median, indicating that the differences between the elite players and the average players are relatively small.

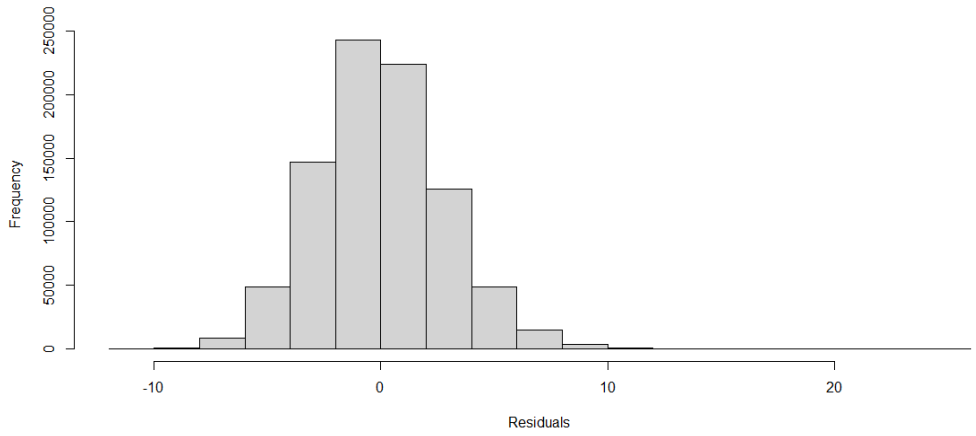


Figure 1: Stage One Residuals Histogram

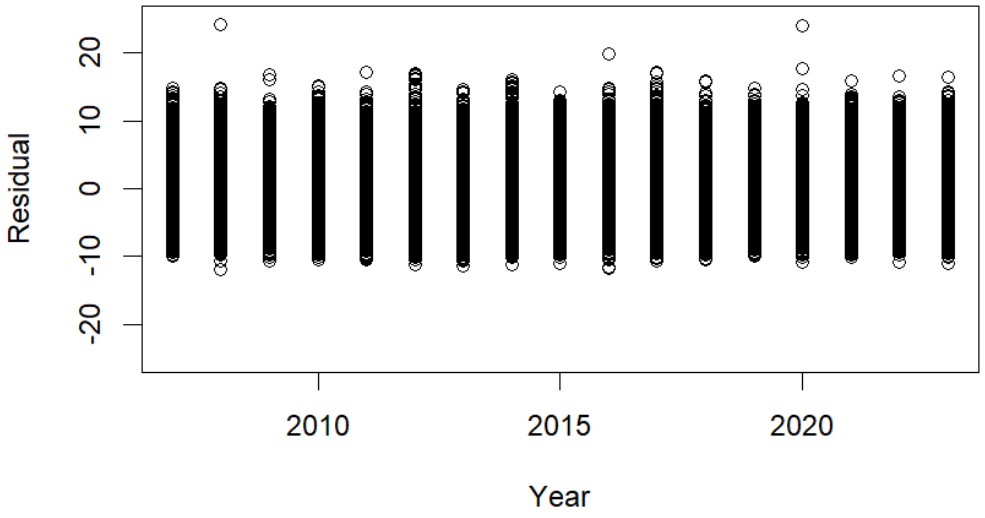


Figure 2: Stage One Residuals by Year

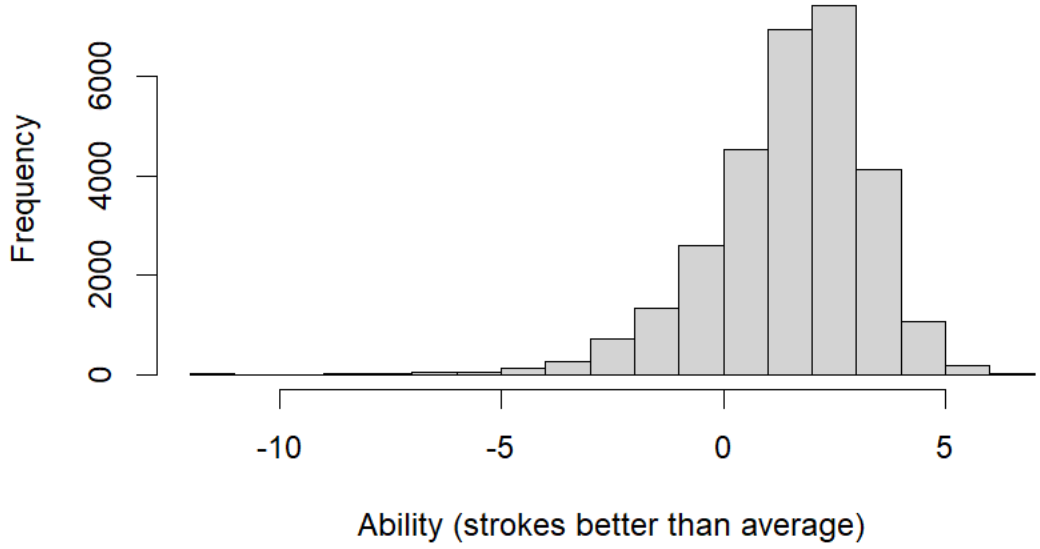


Figure 3: Distribution of Player-Year Ability Estimates

Validity Analysis

To validate the results from the crossed linear mixed effects model in stage one, I performed two validation checks. First, I calculated the correlation between the player-year ability estimates and several other unidimensional measures of golfing ability: a player's ranking in the Official World Golf Rankings (OWGR), total Strokes Gained, the player's ability estimate from the SBSE ("Score-based Skill Estimate") model (Broadie & Rendleman Jr, 2013), and the player's ability estimate from a pooled model that considers only player ability and not player-year ability. All of these estimates are scaled so that higher values indicate greater ability. These correlations (found in Table 6) show that, as expected, the stage one player-year estimates are positively correlated with all of

the other measures of ability. Some of these correlations are in the expected range of magnitudes. However, others are larger than expected. I hypothesized that these correlations would be between 0.6 and 0.8. The results indicate correlations ranging from 0.68 (correlation with OWGR ranking) to 0.979 (correlation with the pooled, age-invariant model). Correlations among the other measures are also presented in Table 6.

The strong correlations provided strong evidence of convergent validity, as the player-year ability estimates produced by the crossed linear mixed effects model are clearly correlated with existing measures of golfing ability. On the other hand, the unexpectedly high correlations that strongly demonstrate convergent validity provide more limited evidence of incremental validity. Because they are quite strongly correlated, one could do almost as well by using one of those existing measures instead of the newly created estimated values.

The second validation check in this stage may mitigate this concern somewhat, fortunately. In this validation check, I used each of the unidimensional measures of golfing ability in Table 6 as predictor variables in separate linear mixed effects regressions. Each was used to replace the player-level and player-year random effects in the original linear mixed effects model (course-related random effects were still included), with each round of golf's score serving as the dependent variable. I then assessed the predictive capability of each of these models using Mean Square Error (MSE).

Table 6: Stage One Validation Correlations

	Stage 1 Estimate	OWGR Ranking	Total Strokes Gained	SBSE Estimate	Age-Invariant Estimate
Stage 1 Estimate	1				
OWGR Ranking	0.680	1			
Total Strokes Gained	0.878	0.729	1		
SBSE Estimate	0.955	0.562	0.658	1	
Age-Invariant Estimate	0.979	0.584	0.659	0.981	1

Results (found in Table 7) provide support for the linear mixed effects model in stage one. The model using the player-year ability estimates from this model outperformed the other models. When performed on the entire dataset, the model using the stage one player-year ability estimates yielded an MSE of 7.7375, which is lower than the other models in this validation check except for the model using total Strokes Gained (MSE=7.48549). However, Strokes Gained are only tracked by the PGA Tour—no other professional golf tour provides the shot-by-shot tracking data needed to calculate Strokes Gained. Thus, the model using the stage one player-year ability estimates was run on a dataset of n=866,147 while the Strokes Gained model was run on a dataset less than a third of the size (n=256,942). The smaller dataset is also likely a non-random subset of the full dataset, containing players of greater ability (because the PGA Tour is the most prestigious professional golf tour) and players who average more rounds of golf (observations) upon which to estimate ability. To counter this, I also ran the model using

the stage one player-year ability estimates as the predictor variable on the restricted dataset of players from the OWGR model. The results indicate that the MSE of the stage one player-year ability estimates did indeed decrease to 7.4248, lower than that of the Strokes Gained model.

Overall, these results provide convincing evidence for the validity of the stage one model. The estimates produced by the model are strongly correlated with the other estimates of ability, and they demonstrated modest improvements in predictive ability compared to the estimates from the other models/measures. Thus, there is evidence to support convergent, predictive, and incremental validity.

Table 7: Predictive Ability of Unidimensional Ability Estimates

	Stage 1 Estimate (n=866,147)	Stage 1 Estimate (n=256,942)	OWGR Ranking	Total Strokes Gained	SBSE Estimate	Age-Invariant Estimate
MSE	7.7375	7.4248	7.9925	7.48549	8.1683	8.1397

Stage Two: Dynamic Measurement Model(s)

Primary Analysis

Marginal Model

To decide which growth trajectory to use for the Dynamic Measurement Model (DMM), I first estimated the five growth trajectories as marginal models (fixed effects only). The model with the best (lowest) BIC and MSE was selected to proceed to the DMM stage. This ended up being the quadratic model, which was the clear winner in terms of both BIC and MSE. The Weibull model was the second-best fitting of the

models, followed by the logistic model, then the exponential model, and, finally, the Michaelis-Menten model as the worst-fitting growth trajectory. The BIC and MSE values can be found in Table 8. The code used to estimate the marginal models and calculate the fit statistics is included as part of the *R* tutorial later in this chapter. Figure 4 shows the fit of the marginal model graphically.

Table 8: Marginal Model Fit Statistics

Model	BIC	MSE
Quadratic	113,312.9	2.7508
Michaelis-Menten	115,145.8	2.9276
Exponential	115,145.7	2.9276
Logistic	114,122.3	2.8265
Weibull	113,811.4	2.7968

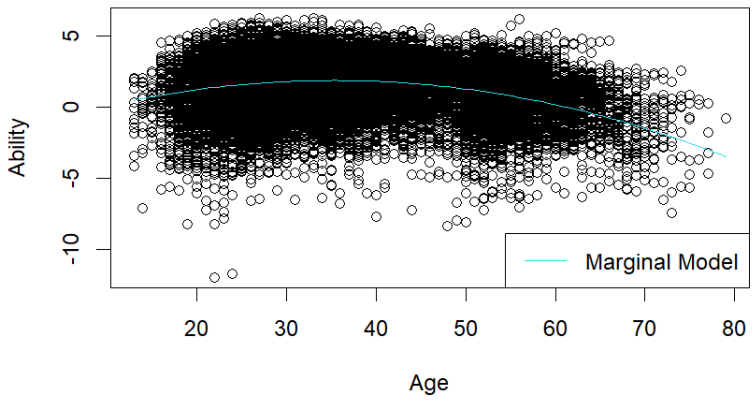


Figure 4: Scatterplot with Marginal Quadratic Model

True DMM Model

I proceeded with the quadratic model as the growth trajectory for the DMM model. This model has three parameters to be estimated and therefore three parameters that can receive random effects. These are a growth rate parameter (B_r), a parameter for the age at which a player reaches his maximum (B_a), and a capacity parameter (B_c). The model with random effects on all three did not converge. Among the quadratic models with two random effects, the combination of random effects with the best fit (lowest BIC) was the model with random effects on B_a and B_c and not on B_r. Table 9 presents the results from this model.

The $B_a = 32.19$ fixed effect estimate implies that, on average, players reach their capacities between the ages of 32 and 33. The random effect on this parameter had a relatively large standard deviation of 7.3686, meaning that, on average, players deviate from the fixed effect value by about seven years. In other words, although the 'average' player reaches his maximum in his early thirties, there is a wide distribution of ages at which players are estimated to reach their maxima.

Table 9: DMM Quadratic Model

Fixed Effects	Estimate	SE	t
Br	-0.001661	0.00003467	-47.91
Ba	32.19	0.2505	128.54
Bc	1.133	0.02489	45.54
Random Effects	Variance	SD	r
Ba	54.2963	7.3686	
Bc	3.0492	1.7462	0.55
Residual	0.1652	0.4064	

The B_c parameter appears to have a smaller standard deviation. However, the unit on this random effect is strokes rather than years. Thus, since the standard deviation of the B_c parameter is 1.7462, players' capacities deviate an average of about 1.75 shots per round from the average (fixed effect) capacity score of 1.133. This means that approximately 95% of professional golfers playing on a major tour are estimated to have capacities between -2.290 and 4.556 strokes above average. Substantively, this seems to be a large difference from player-to-player in terms of capacity scores, which may provide some evidence of the utility of estimating such a quantity.

The correlation between the random effects for B_a and B_c represents the latent correlation between individual golfers' capacities and the age at which they reach their capacities. Because the correlation is positive, golfers with greater capacities are expected to reach their capacities at older ages. This implies that observers may find it difficult to tell from the earliest portions of a player's career how successful that career will be. This latent correlation is moderately strong, implying that it is certainly not determinative, but it is also unlikely to be due simply to sampling variability—it is likely the case that players with the highest capacities do indeed reach their capacities later than other players.

The DMM model can be used to ascertain which players in the dataset have the greatest capacities. One way to conceptualize this is as which player(s) would be most likely to win if all players in the dataset played each other at the peaks of their respective

careers. The 11 players with the highest estimated ability levels are shown in Table 10 below.

Table 10: Top Capacity Estimates According to DMM Model

Player	Country	Capacity Score
Jon Rahm	Spain	5.8337
Rory McIlroy	Northern Ireland	5.6887
Scottie Scheffler	United States	5.4032
Patrick Cantlay	United States	5.3594
Viktor Hovland	Norway	5.3364
Xander Schauffele	United States	5.2990
Justin Thomas	United States	5.2970
Collin Morikawa	United States	5.2534
Dustin Johnson	United States	5.1864
Steve Stricker	United States	5.1541
Tiger Woods	United States	5.0854

Without considering the order of these players, these results are not particularly surprising. All of these players have been ranked in the top three in the world rankings at some point, and seven of them have won multiple major championships. However, the order is potentially surprising, with Tiger Woods having only the 11th highest capacity score in the dataset despite being widely recognized as the best golfer of the current generation of players. Further investigation would be needed to determine whether this is a strength of the model (providing unexpected inferences and conclusions) or a weakness (providing inaccurate conclusions or too much uncertainty).

DMM Tutorial

In this subsection, I present demonstration code for conducting DMM models in *R*. I present the code to run the marginal models for all five growth trajectories, to plot the marginal growth trajectory, to estimate the best-fitting growth trajectory (quadratic) as a true DMM model (including using the marginal models to set starting values for the DMM model), and to calculate BIC and MSE for the DMM model. I also present the full code for a true DMM model in each of the other growth trajectories in Appendix B. Many of these do not achieve convergence on this particular dataset, but the code nonetheless may be useful to future researchers in conducting their own DMM models. To get the code into a format that is ready to be copied into an MS Word document, I use *R Markdown*, which can create PDF, HTML, and MS Word documents with *R* code and output embedded (Baumer & Udwin, 2015).

In this tutorial, I have broken the code up into smaller chunks. These are more digestible for those learning DMM and/or *R* for the first time. This also allows me to provide commentary on the code and results in between the code chunks.

```
###### Quadratic Marginal Model #######
library(foreign)
data.dmm<-read.csv("C:/Users/macwe/OneDrive/Dissertation/Datasets
/player-year.csv", sep="")
data.dmm[c(1:10),]
```

```
##            PLAYER Year      sum1 Age
## 6     A Ilyassyak 2007 -1.884216  42
## 7     A Ilyassyak 2008 -1.211551  43
## 8     A Ilyassyak 2009 -1.320875  44
## 14     A Siddikur 2010  2.453366  26
## 15     A Siddikur 2011  2.552817  27
## 22    A.J. Crouch 2022  2.896804  29
## 23    A.J. Crouch 2023  2.295983  30
## 24    A.J. Elgert 2008  1.837466  26
## 25    A.J. Elgert 2010  2.270091  28
## 27 A.J. McInerney 2017  2.453366  24
```

In this first chunk of code, I used the `library(foreign)` command to load the `foreign` package. This package allows one to load datasets from other (non-*R*) file formats. In this case, I am loading a dataset in a .csv format from a specific file path. If one does not know the file path for the dataset desired, he or she can use the command `read.csv(file=file.choose())` instead, which will then open a file explorer so that he or she can find and select the file on his or her computer. Finally, the last line of code in the chunk above simply displays the first 10 rows of the dataset. The `sum1` variable represents the player-year ability estimates from the stage one model.

```
data.dmm$Age2<-(data.dmm$Age)^2 #creating quadratic term
mod.q<-lm(sum1~Age+Age2, data=data.dmm) #regular linear regressio
```

```
n
summary(mod.q)

##
## Call:
## lm(formula = sum1 ~ Age + Age2, data = data.dmm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3380  -0.9106   0.2056   1.1122   5.5456
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.587e+00  9.831e-02  -16.15   <2e-16 ***
## Age          1.972e-01  5.224e-03   37.74   <2e-16 ***
## Age2        -2.794e-03  6.428e-05  -43.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.659 on 29420 degrees of freedom
## Multiple R-squared:  0.09124,    Adjusted R-squared:  0.09118
## F-statistic:  1477 on 2 and 29420 DF,  p-value: < 2.2e-16
```

The chunk of code above creates a quadratic term and then runs a quadratic marginal model using the `lm` function, which is used for simple linear regressions. This works in this case because the quadratic model can be modeled as a transformation of a linear model. This will not work for the other marginal models, as they cannot readily be converted into transformations of a linear model, but it does work for the quadratic model. The `summary` function then displays the results of the model (coefficients, standard errors, R^2, etc.). Although not strictly necessary in this stage, since we only need to compare the fit statistics (BIC and MSE), it is often advisable to view the results of the model and understand what it is saying.

```
BIC(mod.q)

## [1] 113312.9
```

```
mean((mod.q$residuals)^2)
```

```
## [1] 2.750795
```

These two lines of code calculate the two fit statistics upon which the choice of growth trajectory is being made. The first, of course, calculates the BIC for the quadratic marginal model, while the second provides the mean of the squared residuals, also known as the mean square error (MSE). Because the quadratic model ends up being the best-fitting of the marginal models, I return to it below to plot it and then convert it into a true DMM model. For now, however, I move on to the marginal versions of the other growth trajectories.

```
### Michaelis-Menten
library(drc)
```

```
## Loading required package: MASS
```

```
##
## 'drc' has been loaded.
```

```
## Please cite R and 'drc' if used for a publication,
```

```
## for references type 'citation()' and 'citation('drc')'.
```

```
##
## Attaching package: 'drc'
```

```
## The following objects are masked from 'package:stats':
##
##     gaussian, getInitial
```

```
mod.mm<-drm(sum1~Age,data=data.dmm,fct=drc::MM.3(fixed=c(NA,NA,NA
), names=c("B0","Bc","Bm")))
summary(mod.mm)
```

```
##
## Model fitted: Shifted Michaelis-Menten (3 parms)
##
## Parameter estimates:
##
```

```
##                    Estimate  Std. Error t-value p-value
## B0:(Intercept)  2.4692e+00  3.1998e-02 77.1671  <2e-16 ***
## Bc:(Intercept) -6.5454e+02  5.2998e+02 -1.2350  0.2168
## Bm:(Intercept)  2.4249e+04  1.9615e+04  1.2363  0.2164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error:
##
##  1.711113 (29420 degrees of freedom)
```

Here, I loaded the `drc` package (Ritz et al., 2015) to run the marginal Michaelis-Menten model. Alternatively, one could use the `nls` function in the `stats` package. The `drm` function inside the `drc` package provides a couple of benefits to users, though. First, it does not require the user to code the functional formula for the nonlinear model. Instead, the user can choose between many built-in functions. In this case, the `MM.3` portion of the code tells the `drm` function that it should use the 3-parameter Michaelis-Menten growth trajectory. Second, using the `drc` package means that the user does not need to specify or think about starting values. Selecting good starting values (sometimes called initial values) is a nontrivial difficulty that sometimes arises in nonlinear modeling situations. The choice of starting values can affect whether the model achieves convergence and can affect the parameter estimates. This may be a legitimate barrier to non-technical users wanting to implement DMM models, so using the `drc` package provides a way to avoid this barrier, at least for the marginal model.[13]

[13] The `drm` function in the `drc` package can actually handle random effects as well, so one could theoretically conduct a full DMM model using this package. However, the quadratic growth trajectory is not one that is built-in to the package, so that will not work for the purposes of this study.

The `fixed=c(NA,NA,NA)` portion of the code simply allows all three parameters to be estimated freely by the model rather than being held constant at a specific value. The `sum1~Age` portion of the code tells the `drc` function to use the variable named 'Age' as the independent/time/dosage variable to predict the variable named 'sum1' that serves as the dependent/response variable. Finally, the `data=` argument tells the function which dataset object to use—where to find the variables and data to be used, in other words.

```
BIC(mod.mm)

## [1] 115145.8

mean((residuals(mod.mm))^2)

## [1] 2.927609
```

The code above demonstrates the code to calculate the fit statistics for the Michaelis-Menten model. We can see that both the BIC and the MSE are higher than those for the quadratic model, indicating that the Michaelis-Menten model does not fit as well as the quadratic model did. Indeed, the Michaelis-Menten model ends up being the worst-fitting of all of the growth trajectories. The code for the exponential growth trajectory is shown next.

```
### Exponential
mod.exp<-drm(sum1~Age,data=data.dmm,fct=drc::EXD.3(fixed=c(NA,NA,
NA),names=c("B0","Bc","Br")))
summary(mod.exp)

##
## Model fitted: Shifted exponential decay (3 parms)
##
## Parameter estimates:
##
##                   Estimate  Std. Error t-value p-value
```

```
## B0:(Intercept) -3.3925e+02  2.0499e+02 -1.6550 0.09794 .
## Bc:(Intercept)  2.4711e+00  3.1974e-02 77.2850 < 2e-16 ***
## Br:(Intercept)  1.2589e+04  7.5660e+03  1.6639 0.09614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error:
##
##  1.711111 (29420 degrees of freedom)
```

To model the exponential growth curve, I used the `EXD.3` function in the `drc` package. Other than switching from the `MM.3` function to the `EXD.3` function, the rest of the code is the same. It is worth noting that the creators of the package named this an exponential decay function, but the parameterization can handle both exponential growth and exponential decay.[14] The sign (positive or negative) on the B_r parameter determines this. The code below demonstrates the calculation of BIC and MSE for the exponential model.

```
BIC(mod.exp)

## [1] 115145.7

mean((residuals(mod.exp))^2)

## [1] 2.927604
```

From these results, we can see that the exponential model fits slightly better than the Michaelis-Menten model, but the difference is quite small. The difference for MSE only occurs at the sixth decimal point, and the difference for BIC is only 0.1. Thus, the

[14] The parameterization used by the EXD.3 function is slightly different from the parameterization that I presented early. However, both are exponential models, and both can handle exponential decay and exponential growth depending on the value of the B_r parameter.

quadratic model is still the best-fitting model so far. The code below demonstrates the same process for the logistic model.

```
mod.l<-drm(sum1~Age,data=data.dmm,fct=drc::L.4(fixed=c(NA,NA,NA,NA), names=c("Br","B0","Bc","Bm")))
summary(mod.l)

## Warning in sqrt(diag(varMat)): NaNs produced

##
## Model fitted: Logistic (ED50 as parameter) (4 parms)
##
## Parameter estimates:
##
##                   Estimate  Std. Error t-value   p-value
## Br:(Intercept)  8.5466e-02  2.2797e-03   37.49 < 2.2e-16 ***
## B0:(Intercept) -3.3001e+02         NaN     NaN       NaN
## Bc:(Intercept)  1.8139e+00  1.6022e-02  113.21 < 2.2e-16 ***
## Bm:(Intercept)  1.2426e+02         NaN     NaN       NaN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error:
##
##  1.681344 (29419 degrees of freedom)
```

The logistic model (using the `L.4` function), unlike the ones before it, produces a warning (not the same as an error) that 'NaNs produced.' This warning message is essentially telling us that our model is too complicated for the data and/or not a very good fit for the data. When this happens, some of the standard errors are effectively infinitely large. The `drc` package is handling this by reporting them as nonexistent ('NaN') and sending us a warning message. Interestingly, the other parameters and any fit statistics can still be validly interpreted when this happens. Thus, even though we already have reasons to suspect that this model does not fit particularly well, we can still proceed with calculating BIC and MSE.

```
BIC(mod.l)

## [1] 114122.3

mean((residuals(mod.l))^2)

## Warning in sqrt(diag(varMat)): NaNs produced

## [1] 2.826532
```

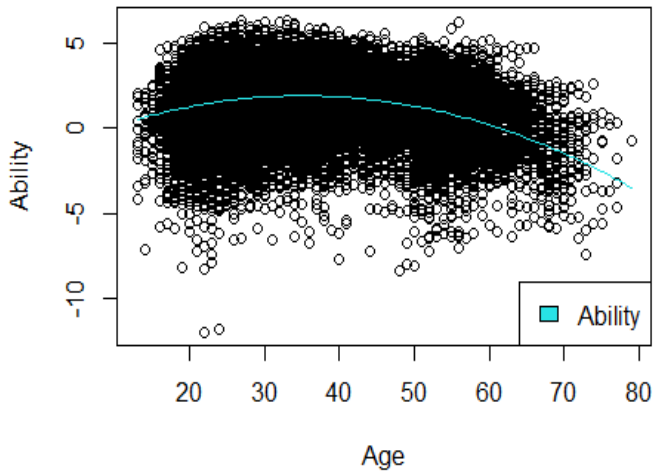
We once again get the same warning message, but the values can still be calculated and interpreted. Perhaps surprisingly, the fit of the logistic model is actually better than those of the Michaelis-Menten and exponential models (lower values of BIC and MSE). The quadratic model remains the best-fitting model tested so far. The code below runs the fifth and final growth trajectory: the Weibull model.

```
mod.w<-drm(sum1~Age,data=data.dmm,fct=drc::W2.4(fixed=c(NA,NA,NA,
NA), names=c("Br","B0","Bc","Bi")))
#summary(mod.w)
BIC(mod.w)

## [1] 113811.4

mean((residuals(mod.w))^2)

## [1] 2.796827
```

The code here is very similar to that of the previous three models using the `drm` function. It uses the `W2.4` function to use one of two different 4-parameter Weibull parameterizations included in the `drc` package. This time, I put a '#' in front of the summary function, which tells *R* not to run this code (commonly known as 'commenting out' that line of code). I also included the BIC and MSE calculations in the same chunk of code. The results indicate that the Weibull model is the best-fitting of the S-shaped and J-shaped growth trajectories that were estimated using the `drc` package, but it still does not fit as well as the quadratic model.

Thus, the quadratic model becomes the selected model for the true DMM model. I proceed below by first plotting the marginal model graphically, as seen earlier in this chapter. I then use the marginal model to identify reasonable starting/initial values for the nonlinear DMM model. This requires manipulating the coefficients a bit to convert them from standard form into vertex form. I then specify the form of the quadratic model to be run in vertex form ($B_r(Age - B_a)^2 + B_c$) so that we get interpretable parameters for capacity and age at capacity.

```
plot(data.dmm$Age,data.dmm$sum1,xlab="Age",ylab="Ability")
curve((-0.002794307*(x^2))+0.197165434*(x)-1.587405578 ,add=T,col
=93)
legend("bottomright",legend="Ability",fill=93)
```




```
mod.q$coefficients

##  (Intercept)          Age         Age2
## -1.587405578  0.197165434 -0.002794307

-1*mod.q$coefficients[2]/(2*mod.q$coefficients[3])
```

```
##       Age
## 35.27984

Ba<- -1*mod.q$coefficients[2]/(2*mod.q$coefficients[3])
(mod.q$coefficients[3]*(Ba^2))+(mod.q$coefficients[2]*Ba)+mod.q$c
oefficients[1]

##     Age2
## 1.890577
```

The coefficient on the quadratic term (-0.002794307) can remain as the starting value for the B_r term. To convert these parameters into an estimate for the B_a parameter, I take the opposite of the linear term and divide it by double the quadratic term: $\frac{-0.197165434}{2*-0.002794307} = 35.27984$. Finally, to get the starting value for the capacity term, I enter the B_a estimate into the Age variable: $\left(-0.002794307 * (35.27984^2)\right) + (0.197165435 * 35.27984) - 1.587405578 = 1.890577$.

```
nform<- ~0+Br*((input-Ba)^2)+Bc
nfun<-deriv(nform, namevec=c("Br","Ba","Bc"), function.arg=c("inp
ut", "Br", "Ba","Bc"))
```

I created the `nform` object to represent the functional form of a quadratic growth function in vertex form: $Ability = B_r(Age - B_a)^2 + B_c$. In this object, "input" is used for the age variable. The `nfun` object uses the `deriv` function to take the equation from `nform` and calculates the partial derivatives for the function. This is a requirement for the `nlmer` function that is used below to actually estimate the DMM model.

```
c.dmm<-nlmerControl(optimizer="bobyqa",tolPwrss=10^-9,optCtrl=lis
t(rhobeg=0.001,rhoend=10^-8,maxfun=40000))
```

In the code above, I set various control settings (also known as hyperparameters) for the DMM estimation. From past experience, these settings provide a good

combination of convergence rates, accuracy, and speed for DMM models. The `lme4` package provides various optimizer options. For linear mixed effects models, the default optimizer is the "bobyqa" optimizer. For generalized linear models, the default in the package is a combination of "Nelder-Mead" and "bobyqa." For true nonlinear models, the default optimizer is "nloptwrap." All of these optimizers are known to be good options; I select "bobyqa" due to my past experience with its ability to balance convergence, accuracy, and speed. The `tolPwrss` hyperparameter tells the optimizer how confident it needs to be in its estimate to 'accept' the solution (i.e. how close to a maximum of the likelihood function it needs to be). Lower values (closer to zero) lead to greater precision but also slower estimation and more convergence problems. The `maxfun` setting tells the optimizer how many iterations to attempt before aborting the estimation: if an acceptable solution is found before reaching this number, then the optimizer will stop anyway. This number represents the number of iterations at which the optimizer will 'give up' if it has not yet found a solution. The `rhobeg` and `rhoend` hyperparameters control how big of "steps" the optimizer takes when it tries a new solution. The `rhobeg` hyperparameter controls the initial step from the starting values, and the `rhoend` hyperparameter controls the size of the step as the optimizer nears the maximum number of iterations.

Some of these settings are specific to the particular optimizer. For example, `rhoend` and `rhobeg` are specific to the 'bobyqa' optimizer; if I had selected a different optimizer, there would have been other settings that could be changed. However, many users may not feel confident changing the values of these hyperparameters. Fortunately,

the `lme4` package provides reasonable defaults for all of these, so someone running a DMM model for the first time could reasonably try the defaults first and then only change settings if convergence is not achieved. Below, I run the true DMM model.

```
model.dmm.q<-nlmer(sum1~nfun(Age,Br,Ba,Bc)~Br+Ba+Bc+(0+Ba+Bc | PL
AYER), data=data.dmm, start = c(Br=-.002794,Ba=35.27984,Bc=1.8906
), control=c.dmm)
summary(model.dmm.q)
```

```
## Nonlinear mixed model fit by maximum likelihood  ['nlmerMod']
## Formula: sum1 ~ nfun(Age, Br, Ba, Bc) ~ Br + Ba + Bc + (0 + Ba
+ Bc |
##     PLAYER)
##    Data: data.dmm
## Control: c.dmm
##
##      AIC      BIC   logLik deviance df.resid
##  54142.8  54200.8 -27064.4  54128.8    29416
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.6713 -0.4918 -0.0064  0.5131  4.9955
##
## Random effects:
##  Groups   Name Variance Std.Dev. Corr
##  PLAYER   Ba   54.2963  7.3686
##           Bc    3.0492  1.7462   0.55
##  Residual       0.1652  0.4064
## Number of obs: 29423, groups:  PLAYER, 5225
##
## Fixed effects:
##      Estimate Std. Error t value
## Br -1.661e-03  3.467e-05  -47.91
## Ba  3.219e+01  2.505e-01  128.54
## Bc  1.133e+00  2.489e-02   45.53
##
## Correlation of Fixed Effects:
##    Br     Ba
## Ba -0.212
## Bc -0.162  0.241
```

Using the `nlmer` function in the `lme4` package, I use the `sum1 ~ nfun(Age , Br , Ba , Bc) ~ Br + Ba + Bc + (0 + Ba + Bc | PLAYER)` expression to specify the quadratic functional form. The `nfun` function was already specified earlier as being an *R* object with a quadratic form along with partial derivates calculated by *R*. When I created the `nfun` function, I specified that `function.arg=c("input" , "Br" , "Ba" , "Bc" ))`. This tells it that these are the inputs or parameters that are used to generate the outcome. Thus, the `nfun(Age,Br,Ba,Bc)` argument now mimics that format, replacing "input" with the name of the input variable in the data.dmm dataset ("Age"). Because the other values are parameters to be estimated rather than actual inputs from the dataset, they do not change names. The next part of the expression details the fixed and random effects to be estimated. The `Br + Ba + Bc` portion outside of parentheses indicates the fixed effects to be estimated, while the `(0 + Ba + Bc | PLAYER)` portion indicates that random effects should be estimated for the B_a and B_c parameters (and that the grouping variable within which they are varying is defined by the PLAYER variable in the dataset). The zero indicates that no additional intercept term should be estimated. The rest of the code specifies the starting values, control settings, and name of the dataset. The code below calculates the BIC and MSE of this model that are used below during validation and demonstrates how to extract the individual-level random effect values from the model.

```
BIC(model.dmm.q)
## [1] 54200.8
```

```
mean(residuals(model.dmm.q)^2)
```

```
## [1] 0.1307608
```

Finally, the code below is used to extract the individual-level random effects. It then adds the individual-level capacity random effect estimates to the capacity fixed effect in order to obtain the overall capacity estimate for each player. It then shows the first 10 rows of the resulting dataset that includes total capacity estimates.

```
r.d<-ranef(model.dmm.q)
r.d2<-r.d$PLAYER
r.d2[c(1:10),] #shows the first 10 rows
```

```
##                          Ba         Bc
## A Ilyassyak     -5.7433766 -2.1138154
## A Siddikur       2.8573574  1.4499253
## A.J. Crouch      2.8298545  1.4714604
## A.J. Elgert      2.2864904  0.9884012
## A.J. McInerney   2.7506917  1.4779002
## Aadil Bedi      -2.5705819 -1.4461048
## Aaron Baddeley   1.4427813  2.5742180
## Aaron Black     -5.4168317 -2.2644922
## Aaron Cockerill  3.9981883  1.8128985
## Aaron Goldberg   0.8008481  1.4943871
```

```
r.d2$Capacity<-fixef(model.dmm.q)[3]+r.d2$Bc
r.d2[c(1:10),]
```

```
##                          Ba         Bc   Capacity
## A Ilyassyak     -5.7433766 -2.1138154 -0.9806722
## A Siddikur       2.8573574  1.4499253  2.5830685
## A.J. Crouch      2.8298545  1.4714604  2.6046035
## A.J. Elgert      2.2864904  0.9884012  2.1215444
## A.J. McInerney   2.7506917  1.4779002  2.6110434
## Aadil Bedi      -2.5705819 -1.4461048 -0.3129616
## Aaron Baddeley   1.4427813  2.5742180  3.7073612
## Aaron Black     -5.4168317 -2.2644922 -1.1313490
## Aaron Cockerill  3.9981883  1.8128985  2.9460417
## Aaron Goldberg   0.8008481  1.4943871  2.6275303
```

Validity Analysis

The validation analysis consists of three components: estimating restricted-sample DMM models and comparing the estimates from these to the estimates from the full model, estimating a linear growth trajectory with random effects and comparing the fit of this model to that of the DMM model, and comparing the predictive ability of single-timepoint estimates to the predictive ability of the capacity estimates from the DMM model. The results generally confirmed the benefits of the DMM model. The estimates from the restricted models show very strong correlations with those from the full model, indicating strong reliability for the model: the estimates are not dependent on particular timepoints, and they are not dependent on having late-career datapoints and scores. The DMM capacity estimates, both from the full model and from early-career restricted models, were far more effective than OWGR and total Strokes Gained at predicting players' highest year ability scores. Conversely, the DMM capacity scores show only relatively small predictive improvements over single-timepoint estimates estimated by the linear mixed effects model in stage one.

Correlations with Restricted Samples

In this portion of the DMM validation, I calculated the correlations between the full-model capacity scores and capacity scores from two types of restricted model. First, I ran the DMM model on two age-restricted datasets: one in which only player-year observations under the age of 30 were included and another in which only player-year observations under the age of 25 were included. Second, I used a modified version of 10-fold cross-validation. I divided the dataset into 10 different, randomly selected subsets,

ran the DMM model on each subset, and calculated the correlation between the capacity scores from the full model and the capacity scores from the subsets.

The capacity scores from the age-restricted models show very strong correlations with the full-model capacity scores. When the dataset was restricted to observations when players were under 30 years old, the capacity scores correlated with the full-model capacity scores at $r = 0.9934$. When the dataset was further restricted to only those player-year observations under age 25, the strength of the correlation decreased only slightly to $r = 0.9876$. The strength of these correlations is remarkable, providing strong evidence for the reliability of the capacity scores—they are not dependent on particular timepoints in general or on having late-career timepoints specifically.

The modified 10-fold cross-validation provided similarly strong evidence for the reliability of the capacity scores. Each randomly selected subset of the dataset produced capacity scores that correlated with the full-model capacity scores between $r = 0.98$ and $r = 0.99$. The lowest correlation for any of the 10 datasets was $r = 0.9807$ and the greatest was $r = 0.9824$. Thus, not only does this support the reliability and validity of the results of the DMM model in this study, but it also may provide some evidence that capacity scores from DMM models more broadly can be considered reliable.

Comparison to Linear Growth

In this portion of the validation, I compared the fit of the quadratic DMM model to that of a baseline linear growth model. In this portion of the validation, the results were not quite so clearly positive. The linear model did not converge, which should provide some skepticism about its estimates. Nonetheless, the non-converged linear model could

still generate fit statistics The quadratic DMM model produced a lower BIC than the linear model, though the two BIC values were similar (see Table 11). The MSE, on the other hand, produced the opposite result: the mean square error was similar for the two models, but the linear model actually had the slightly lower MSE, indicating better fit.

This is somewhat confusing, as a linear model should be nested within a quadratic model, making it theoretically impossible for the linear model to have a better MSE than the quadratic model. This result is possibly due in part to the non-convergence of the linear model. More likely, however, is that the DMM parameterization of the quadratic model in vertex form was not performing as well as a version in standard form would have.[15] Accordingly, I confirmed this by estimating a quadratic mixed effects growth model in standard form. Similarly to the DMM model, the model did not converge when it had random effects on all three parameters. A model with two random effects (one on the linear term and one on the constant term) does indeed show that it outperforms the linear model.

Table 11: Linear vs Quadratic Fit

	BIC	MSE
DMM Quadratic Model	54,200.8	0.1308
Standard Form Quadratic Model	53,460.5	0.1183
Linear Model	54,275.4	0.1195

[15] Without random effects, they would be equivalent. However, with random effects added, one could provide better fit than the other.

The support for the validity of the quadratic DMM model is more tepid in this stage of the validation. Its fit does not provide much, if any, improvement over a linear growth model. Nonetheless, because the linear model did not converge in *R* and because the DMM model provides the benefit of the readily interpretable capacity parameter, it likely still provides some benefit above the linear model in terms of convergence and parameter desirability. Similarly, because all three models fit similarly, it may be beneficial to select the DMM parameterization of the quadratic model over the standard form quadratic model. This may sacrifice a small improvement in fit (compared to the standard form quadratic), but it has the benefit of the interpretable capacity parameter.

Comparison to Single-Timepoint Estimates of Ability

The final portion of the validation compares the predictive ability of the DMM model to the predictive abilities of various single-timepoint measures of ability. I tested the capacity estimates from the full DMM model, the capacity estimates from age-restricted DMM models, the single-timepoint ability estimates from the linear measurement model in stage one of this study, single-timepoint OWGR rankings, and single-timepoint Strokes Gained values. For any given age, the OWGR rankings, the Strokes Gained values, and the stage one latent ability estimates simply used the value at that age for the particular golfer. For the restricted-age DMM model, the model was estimated on all datapoints for that age and younger (e.g. if I was testing the predictive ability at age 27, all players at all ages less than and equal to 27 would be included in a dataset that was then used to estimate a DMM model and calculate a capacity score). For each measure of ability at each age, I ran a simple linear regression with a player's

maximum estimated ability from stage one as the dependent variable. The R^2 for each regression represents the predictive power of the single-timepoint measure of ability to forecast future ability. Table 12 displays the results.

The latent measurements of ability and capacity produced in this study clearly outperformed the existing single-timepoint measures of golfing ability at forecasting future maximum player-level ability values. The full-model DMM capacity estimates, the restricted-model DMM capacity estimates, and the single-timepoint latent ability estimates from stage one all were very effective predictors of players' maximum ability levels: they all had R^2 values greater than 0.9, making them very effective predictors. By comparison, OWGR rankings and total Stroked Gained were less effective predictors of a player's maximum future ability level.

However, among the latent variable measures of ability, the DMM models provided only marginal improvements in forecasting accuracy compared to the single-timepoint latent ability estimates from stage one. Interestingly, some of the age-restricted DMM models actually outperformed the full-model DMM model. This is possibly due to selection effects: the only players in the dataset at particularly young ages are the ones who end up having more successful, more predictable careers. Regardless, this portion of the validation provides strong evidence for the benefits of latent variable models over observed-variable models, while there is some evidence for DMM models (or longitudinal models more broadly) over single-timepoint latent measurements, but this evidence is not as strong.

Table 12: Predictive Ability (R^2) of DMM vs Single-Timepoint Estimates of Ability

Age	DMM Full Dataset	***DMM*** Restricted-Age	Stage One Model	OWGR	Total SG
Full Dataset	0.9544				
30		0.9552	0.918	0.464	0.4775
29		0.9537	0.9204	0.4745	0.4814
28		0.9520	0.9184	0.429	0.4992
27		0.9522	0.9252	0.4322	0.5811
26		0.9459	0.9257	0.4233	0.6178
25		0.9422	0.9174	0.3661	0.4167
24		0.9393	0.9185	0.348	0.4925
23		Did not converge	0.9315	0.3117	0.5684
22		Did not converge	0.9464	0.2161	0.4417
21		Did not converge	0.9500	0.1855	0.4066
20		Did not converge	0.9518	0.1597	n too small
19		Did not converge	0.953	0.0006	n too small
18		Did not converge	0.9414	0.1493	n too small

Conclusions

The results broadly indicate that modeling the latent abilities and latent capacities of professional golfers was successful. This study had four research objectives: to summarize and synthesize the literature on the quantitative measurement of golfing abilities, to estimate the abilities of professional golfers as a latent variable, to assess the longitudinal shape of ability growth over time, and to measure the latent capacities of

professional golfers. The literature was synthesized in Chapter Two. Stage one of this chapter accomplished the second objective by using a crossed linear mixed effects model. The resulting player-year ability estimates from this model were shown to be positively correlated with existing measures of golfing ability while still being better able to predict players' scores than those existing measures.

Stage two of this chapter addressed the third and fourth research objectives. The quadratic growth model fit better than any of the alternative DMM growth trajectories according to both measures of model fit. The quadratic DMM model successfully estimated player capacities. These capacity estimates were shown to be highly reliable: they were consistent and highly correlated across various subsets of the data, and early-career capacity estimates were shown to be nearly perfectly correlated with capacity estimates from all data points. These capacity estimates (both from the full model and from early-career age-restricted models) were also much more effective at forecasting future ability than existing single-timepoint observed-variable measures golfing ability. However, these estimates provided only small improvements over latent single-timepoint estimates from the measurement model in stage one of this study, and the quadratic DMM model did not provide much improvement in fit from a baseline linear growth model.

Although the DMM model from stage two provides only moderate improvements over other latent variable models and approaches to measuring capacity, it still provides significant improvements over any existing observed-variable approach to measurement. Thus, this study has provided evidence that the statistical benefits of treating golfing

ability as a latent variable rather than as an observed variable are clear and unambiguous. The estimates from both latent variable models (the crossed linear mixed effects model in stage one and the quadratic DMM model in stage two) consistently outperformed any of the unidimensional observed variable measures of golfing ability, including Strokes Gained, the current state-of-the-art measure of golfing ability. The latent variable approaches presented here have shown high reliability, convergent validity, predictive validity, and incremental validity; the results from those models are both trustworthy and improve upon existing methods and measures.

Chapter Five: Discussion

In this study, I have presented evidence for treating golfing ability as a latent variable rather than as an observed variable, synthesized the existing literature on the measurement of golfing ability and on mixed effects models used as measurement models, developed novel latent variable statistical models to measure golfing ability at individual timepoints and to measure a player's capacity for future ability growth, and validated these models. The latent variable approaches developed and presented in this study unambiguously outperformed all unidimensional observed variable approaches to the measurement of golfing ability, including Strokes Gained. As Strokes Gained is widely accepted as the current state-of-the-art method for measuring golfing ability, this was a high bar to clear. In this chapter, I proceed by discussing the limitations of the study, the contributions that the study has made to two different audiences, and recommended directions for future research.

Limitations

Unidimensionality

Despite its success in demonstrating the improvement in measurement that can be made by switching to latent variable methods to measure golfing ability (which is inherently a latent ability), the study does still have several limitations. First, and perhaps most notably, the measures of golfing ability and capacity that I have created in this study

are both still unidimensional measures of golfing ability. They implicitly assume that golfing ability is a single monolithic ability that can be measured using a single number for any given time point. This is a strong assumption, and it remains untested.

It is quite plausible that golfing ability might be better conceptualized as multidimensional, with players having different combinations of skills such as driving ability, putting ability, approach shot ability, etc. Previous research has pursued this path using observed variable approaches to the measurement of golfing ability, but such studies have at least two weaknesses. First, they have used observed variable approaches, and the research in my study has shown that observed variable approaches—even sophisticated ones—are not likely to be as effective as latent variable approaches. Second, instead of ascertaining the dimensionality of golfing ability inductively or by hypothesizing a certain structure or dimensionality of golfing ability and then testing it, previous multidimensional studies of golfing ability have assumed a given factor structure and never tested it. These assumed structures are based on subject-matter understandings of golf as a sport, and they represent logical and reasonable theories, but they remain just assumptions that have not been tested.

As such, the extant research on golfing ability as a multidimensional construct has not been particularly sophisticated statistically. No known study has yet combined a true latent variable approach to the measurement of golfing ability with a multidimensional mindset. Nonetheless, it is entirely possible that this would provide the best description of reality and therefore be the most effective way to model golfing ability. For now, though, it remains solely a direction for future research.

Fortunately, the assumption that golfing ability is unidimensional, even if inaccurate, seems intellectually more defensible than assuming a different factor structure without testing the accuracy of such a structure. In some situations, we must first define our latent construct and decide that it does exist before proceeding to understand its dimensionality. Just as it would be difficulty to measure the dimensionality of intelligence if we do not already know or agree that the construct of 'intelligence' exists, it would potentially be difficult to measure golfing ability as a latent multidimensional construct if we did not first establish that latent golfing ability exists. This study, despite the potential limitation of assuming that golfing ability is unidimensional, does provide evidence that latent golfing ability exists and can be measured. Assessing its factor structure will be a direction for future research.

Arbitrary Definition of "Average"

One weakness or limitation of existing studies on golfing ability is that many of them must arbitrarily define an "average" golfer as a reference category to which others can then be compared. This is sometimes accomplished quantitatively (e.g. Strokes Gained) and sometimes it is done by human definition of expectations (e.g. handicap rating systems and comparisons to "par"). The problem with this is that it makes the quantitative value placed on each individual's ability level seem less substantively meaningful, less substantively interpretable, and less transferable across populations. For example, if a professional golfer on the PGA Tour plays a tournament and has negative two Strokes Gained relative to the field of other professionals, and I play a tournament at my local golf course and have positive three Strokes Gained relative to the other players,

how do I compare my positive three Strokes Gained with the professional's negative two Strokes Gained? They are in relation to different populations.

It is easy enough to fix this problem and simply define a single point on the ability spectrum to which everyone would be compared. For example, if an amateur calculates his or her Strokes Gained for a round and finds that he or she has negative 18 Strokes Gained relative to the average professional golfer, that would solve the problem of comparing to different populations. However, it would still be an arbitrary choice for where to anchor that reference point.

The measures of golfing ability that I have created in this study suffer from this same arbitrariness. Because my population was professional golfers on major tours (PGA Tour, PGA Tour Champions, DP World Tour, Korn Ferry Tour, and LIV Golf), the reference point to which all golfers in the dataset were compared became the average ability level among golfers on major tours. This is arbitrary, though. If I had included some of the smaller golf tours, such as the Canadian Tour or the Japanese Tour, the average ability level in the dataset would have dropped, which would then have raised the estimated ability level for those already in the current dataset. Thus, the player-level ability levels estimated by both the stage one and stage two models are only partially meaningful in an absolute sense.[16]

Fortunately, though, they remain meaningful and useful in a relative sense, regardless of the interpretation in an absolute sense. This is common in quantitative

[16] More formally, this means that we could think of the ability estimates as operating on an interval scale rather than on a ratio scale.

measurement—in many cases, ability estimates are simply on a standard deviation scale with no substantively meaningful units. An ability of 0.98 would simply mean 0.98 standard deviations above average. The models in this study do at least have meaningful units (golf strokes) on the ability variables, but these values, as with many measurement models, are dependent on the definition of the population at hand. For substantive users of these measures, this may be less satisfying, even if it is not particularly problematic statistically.

Equally Informative Golf Courses

Like a Rasch model in the context of Item Response Theory, in which each item on a test is assumed to be equally discriminating in relation to the underlying latent ability, the models used in this study inherently assume that a round of golf played on one golf course provides the same amount of information about a player's ability as a round of golf played on another golf course. It is theoretically possible, however, that some golf courses do a better job of measuring latent golfing ability than do other golf courses. Thus, the use of this implicit statistical assumption may lead to scores from less discriminating courses having too great of an effect on players' estimated ability levels while scores from more discriminating courses receive less weight than they ideally should.

There are benefits to this more Rasch-like approach, however, in terms of interpretability and computing time. The stage one model in this study already took almost 24 hours to run on a computer with a powerful Intel Core i7 Processor. Adding additional random effects for course-level discriminations/weights/loadings would almost

certainly slow down the estimation of the model further, and it would make the functional form of the model and the ability estimates less readily interpretable. The stage one model in its current form can be written in a simple additive linear format with two of the estimated parameters combining to create the player-year ability estimates. Allowing the discriminations/weights/loadings to vary would complicate this functional form and would make it more difficult to understand the model's parameters.

Nonetheless, it is possible that such a model would provide improved fit and therefore more accurate player ability estimates. Fortunately, the stage one model performed quite well, as its estimates proved to be better predictors of golf scores than other existing unidimensional measures of golfing ability. Thus, the assumption of equal discriminations, while potentially sub-optimal conceptually, does not appear to be causing significant problems from a practical standpoint.

Two-Stage Dynamic Measurement Model

The original conceptualization of DMM models involved the simultaneous estimation of single-timepoint estimates and capacity scores in a single model. DMM models are measurement models, so the estimated ability level at any point on the growth curve can be considered the individual's estimated ability level at that point in time. It is therefore redundant, theoretically, to first estimate individuals' ability levels at single timepoints from one model and then use these estimates in a separate model to estimate the longitudinal growth trajectory. Not only is it redundant, but it also ignores uncertainty in the stage one estimates by treating them as observed variables during the second stage. A player-year ability estimate from stage one that is based on two rounds of golf (or two

test items, etc.) will be treated with equal confidence as a player-year ability estimate from stage one that is based on one hundred rounds of golf (or one hundred test items, etc.). Because we have (and, by extension, the model has) more information about the second player-year, we should expect it to have a greater effect on the shape of the longitudinal growth trajectory.

With a model in which the ability estimates and the growth trajectory are estimated simultaneously, this would happen. However, in the two-stage model, these two player-year observations will be treated as equally informative during stage two. This may lead to inefficient estimates from the stage two DMM model. Thus, in an ideal world, one would use the original conceptualization of DMM models as measuring timepoints and capacity simultaneously instead of estimating two separate models.

However, two-stage models have become the norm in DMM research for practical reasons, even among the original creators of the DMM modeling framework. The simple reason for this is that it takes a significant amount of time and computing power to estimate the DMM model in a single model as the creators of DMM originally envisioned. The crossed linear mixed effects model used in stage one of this study took almost 24 hours to converge. Building a nonlinear growth trajectory overlay onto this in the same model would likely increase computing time significantly. Thus, if one needs the ability estimates in a timely manner (or if one needs his or her computer to perform other tasks), such a model may be impractical, and the two-stage model may provide a more practical alternative. Nonetheless, this is still a limitation of the study, as it is still theoretically a sub-optimal modeling approach.

Golf Scores are Noisy Measures of Golfing Ability

As other studies and researchers have noted, players' results from individual golf holes, individual rounds of golf, and individual tournaments are not perfect indicators of golfing ability. The best player does not win every tournament, does not always have the lowest score in a round, and does not always perform the best on every hole. During the years included in the dataset in this study, the most consecutive tournament wins by a single player was five by Tiger Woods in 2007-2008. Even the best players have bad days, lose tournaments, etc.

The crossed linear mixed effects model from stage one was very successful at predicting scores better than existing measures of player ability. However, the model still had a residual standard deviation of over 2.8 shots, indicating that, even after accounting for the player's ability level, the course's difficulty on that particular day, etc., a player's score in a particular round would be still be expected to deviate from its "true" score by over 2.8 shots on average. Because of this, seeing a single round of golf in which one player shoots 75 and another shoots 69 tells us almost nothing, statistically, about which of those two is a better golfer. This is a situation where humans might be tempted to over-interpret such a result, but the statistical model is better suited to assessing what this actually tells us.

As such, the models presented in this study inherently are large-n methods. They are not going to do a particularly good job of assessing an ability level or a capacity for players who only have a few rounds in the dataset. They will be much more effective and accurate for players with hundreds of rounds in the dataset.

In some cases, researchers interpret a large residual variance in a mixed effects model as meaning that the model is not doing a very effective job of fitting the data. However, Carsey & Harden (2013) differentiate between fundamental uncertainty and estimation uncertainty. In any statistical model that is used for statistical inference, there is going to be some uncertainty because the model is using sample data to draw inferences about a population. This uncertainty is known as estimation uncertainty, and it is unavoidable without having the entire population in the dataset. Because estimation uncertainty is unavoidable, we often interpret residual errors as denoting estimation uncertainty—if we only built a better, more sophisticated model, we could continue to reduce the residual variance term.

This may true in many cases, but there is also another source of error that contributes to the residual variance or standard deviation term: fundamental uncertainty. Some data-generating processes in the real world may have some randomness built into them. In the golf context, three players could all hit their drives into the same tree. One of them could get a favorable bounce off the tree into the fairway, another could get a neutral bounce off the tree into the rough, and the third could find that his ball gets stuck in the tree and is unplayable. These differences in outcome are not based on any skill or ability—it is not the case that any of them tried to hit the tree, nor is it the case that the differences in outcome come from which player is better at ricocheting the ball off of the tree. Since these differences in position are likely to affect the players' scores on that golf hole, a random process will have affected the players' scores.

Such random processes exist in life and in golf. To the extent that the residual standard deviation is capturing these fundamental uncertainties that occur in the data generating process of playing golf rather than capturing estimation uncertainty, the large residuals may not be entirely concerning conceptually. From a measurement perspective, however, they are still not particularly desirable, and they represent a limitation to the study as they make it more difficult to accurately measure ability and require larger numbers of observations to obtain reliable estimates.

Barriers to DMM

Although I provided a tutorial for conducting Dynamic Measurement Models in *R*, it was likely not as simple as some users would like it to be, and there are still relatively high barriers to conducting DMM models in either *SAS* or in *R*. Both software require writing code/commands, providing starting values, understanding how to write the equation form of the growth trajectory into the command, etc. Note that these second two barriers are different from linear models and generalized linear models: even in *R*, the user does not need to write the equation of the model into the command, nor does the user need to specify starting values for linear models or generalized linear models. Thus, to conduct DMM models, one must already possess or be willing to learn at least a basic understanding of programming (writing code) and the mathematical formula for a specific growth trajectory. For some researchers and practitioners, these may constitute enough of a reason to select a different type of model.

Some Tournaments Excluded from Dataset

A few tournaments were excluded from the dataset due to having nontraditional scoring. For example, match play events could not be included, because there is no player-round score that results from these tournaments. Instead, one player simply beats his opponent for the day and then moves on to the next opponent. If they played all 18 holes, it might be possible to calculate a score equivalent for each player. However, in many cases, they will not finish the round once one player has already won.[17] Similarly, a few tournaments have used a scoring system known as a Stableford system which uses points rather that strokes and rewards good scores more than it punishes bad ones (effectively a nonlinear scoring system). Finally, there are occasional team events in which players alternate shots with a partner. Because these cannot be compared on the same scale as regular golf scores, these three categories of events were excluded from the dataset.

As long as players' performance in these events is 'caused' by the same latent underlying ability as the performance in stroke play events, then the omission of these tournaments should not cause bias. Nonetheless, this is an untested assumption. Additionally, more observations are always better, especially when trying to create accurate measurements in the context of high residual variance, so losing any observations is generally not desirable.

DMM Model's Modest Improvements

[17] Match play tournaments typically track how many holes a player wins rather than overall number of strokes for the entire round.

While this study overall demonstrated clear evidence that measuring golfing ability as a latent ability provides an improvement over existing measures of golfing ability, the quadratic DMM model in stage two provided only modest improvements in fit over a baseline linear growth model and only modest improvements in predictive ability over latent single-timepoint estimates from stage one. The stage two model runs much more quickly than the stage one model (measured in seconds rather than hours), so these modest improvements may still be justifiable and worth the computational effort. The question is not whether the quadratic model was the right choice for this study given the alternatives tested. Nonetheless, the relatively modest improvements may provide some doubt about the 'true' shape of longitudinal ability growth over time for professional golfers on major tours.

Contributions

This study makes contributions to two distinct audiences and bodies of literature: the literature on the measurement of golfing abilities and the literature on dynamic measurement modelling. There is certainly some overlap between these contributions, but it makes sense to organize the contributions by the audience that will be most interested in each. I proceed first with the contributions to golf measurement and then follow them with the contributions to the field of DMM research.

Golf Measurement

Applicability of Latent Variable Statistical Models

Perhaps most importantly, this study has demonstrated the efficacy and applicability of latent variable statistical models to the study of golf specifically and to

the study of sports more generally. There have been a handful of previous studies to do this, but the vast majority of sports statistics, even within advanced analytics, still rely on observed variable data and models. This study has demonstrated that a latent variable approach can outperform even the most advanced observed variable approaches.

In Stage one, the estimates of latent golfing ability from the linear mixed effects model showed a strong correlation with total Strokes Gained ($R^2 = 0.878$), providing convergent validity for the estimates of latent ability, but the estimates from the model proved to be better predictors (lower MSE) of players' scores than did total Strokes Gained (once they were run on the same set of data). Similarly, the stage one single-timepoint estimates and the stage two capacity estimates both proved to be much better predictors of players' future maximum ability levels than did total Strokes Gained. This is despite the fact that total Strokes Gained consistently showed that it was more effective than the Official World Golf Rankings at prediction and forecasting.

Strokes Gained is widely accepted by scholars, television commentators, and players as the most modern, advanced, and state-of-the-art way of measuring golfing ability. It is so ubiquitous as a statistic in the professional golf industry that players are asked about it during interviews, commentators mention it during broadcasts, and the website for the Official World Golf Rankings reports Strokes Gained along with players' official rankings. This is fairly impressive for a statistical measure of ability that was first published as a book chapter in 2008 (not even called "Strokes Gained" yet) and then more widely disseminated in a 2012 journal article that used the term Strokes Gained for the first time.

Strokes Gained arose as a *solution* to the problem of previous observed variable statistics being inadequate. It is still based on observed variable logic, but it is a much more sophisticated measure of golfing ability. Thus, outperforming Strokes Gained is a significant achievement and provides strong evidence for the utility of latent variable statistical models (like the ones that I have presented in this study) to the study of golf and to the study of sports more broadly.

The Models in this Study do not Require Shot-Level Data

Unlike Strokes Gained and other advanced statistical measures of golfing ability, the methods presented in this study do not require shot-level data. The collection of shot-level data is fairly onerous—so much so that only one professional tour tracks data granularly enough to calculate Strokes Gained for its players and fans. Other professional tours do not have this level of data; college teams, high school teams, and recreational golfers are also unlikely to have this level of data. Because of this, many who would be interested in tracking ability levels, via Strokes Gained or similar measures, for themselves or for others are unable to do so.

The measures of golfing ability in this study not only are potentially more accurate than Strokes Gained, but they also require significantly lower levels of data collection. Instead of collecting four pieces of data for every shot as Strokes Gained would require, only a single number is needed for the entire round: the score for that round. Most players are naturally collecting their score for the round anyway, so there is likely no new data collection needed. This method provides a much more accessible way for recreational golfers, amateur golfers, and lower-level professional golfers to track

their ability levels. Because this study shows that it can outperform Strokes Gained anyway, these golfers need not even sacrifice accuracy.

Goes Beyond Single-Timepoint Measures of Ability

Existing measures of golfing ability inherently measure ability at a single point in time. No existing measure was designed to forecast future ability growth. The DMM model from stage two of this study changes this. In this stage, I tested several candidate growth trajectories, selected the best-fitting trajectory, and then implemented a model than can be used to forecast player abilities at a specific point in the future, and, perhaps more interestingly, can be used to estimate each player's *capacity*, or the player-level maximum ability that is modeled to occur in the future or the past.

Especially for those in the golf industry who are using measures of golfing ability to make decisions about the future, the implementation of a capacity estimate may be of particular interest. For example, if a college coach is trying to decide which players to recruit, he or she would be able to use the DMM model to forecast each player's capacity rather than solely the player's current ability. Alternatively, if players' capacities tend to occur in their early thirties, maybe the college coach cares less about the capacity estimate *and* less about the current timepoint estimate at age 16. Instead, maybe the college coach would use the model to estimate each player's forecasted ability at age 21 to decide whom to recruit. Similarly, sponsors and tournament organizers may also be particularly interested in players' capacity estimates when they decide which players to sponsor or admit to their tournaments—even if a player has lower ability at the current time, they may want to plan for the future by building relationships with those players

who will develop greater ability in the future. Finally, just as DMM can be used in the education context to reduce the influence of demographic factors on estimates of future ability and identify students with high capacities even if they have lower current ability, a DMM model in the context of golf may be able to identify young players from underserved countries and communities who may not have as high of current ability levels but who show promise for future ability growth. These players could then be targeted with scholarships and opportunities for instruction to help grow their skillset for the future.

Provides Valid Comparisons across Professional Tours

The latent variable models presented in this study provide the ability to simultaneously estimate different players' abilities on the same scale even if they have never played in the same tournaments or on the same courses as each other. The model accomplishes this by using cross-pollination between tours and courses to link other players and courses. This provides a statistically valid means of comparing the ability levels of golfers across time and space.

The OWGR try to do this as well, as the rankings rank players from various tours into the same ranking system. However, previous research has shown that the OWGR are biased against the best players (Broadie & Rendleman Jr, 2013). Additionally, the OWGR (like many ranking systems in sports) incentivizes players to play more tournaments. In the case of the OWGR, players' rankings are punished if they do not play at least 20 events per year. This makes sense from a financial perspective, but it makes less sense from a measurement perspective: the player has the same ability level whether

he plays 18 events or 21 events in a year. Thus, the rankings may be more accurately thought of as a measure of recent accomplishment rather than as a measure of ability.

Furthermore, the OWGR strongly favor players who win tournaments over others who finish near the leader. The winner receives 100% of the tournament's "value," while the second-place finisher receives only 60% of the tournament's "value." However, from a statistical perspective, a player winning a tournament and a player finishing one shot behind the leader have displayed very similar levels of ability, so the OWGR's emphasis on winning alone is likely due to some other goal than solely the measurement of golfing ability. Thus, providing a more statistically-based and statistically-valid way to compare player ability levels across tours should be an improvement over currently available options like the OWGR.

Improvement to Course and Handicap Rating Systems

The existing handicap rating system requires expert evaluators to visit each golf course periodically and assess its difficulty and discrimination ("course rating" and "slope," respectively) based on their semi-subjective assessments of how it would be played by players of varying ability levels. Then, when players play the course, their performance is judged against the baseline expectation that the evaluators created, earning the player a handicap—how many strokes above par the given player would be expected to score on an average day at an average course. If the evaluators misjudge a certain course, this then can create bias in individual players' handicaps. For example, if the evaluators judge a course to be too easy, then a certain score will appear to display

less ability than would actually be required to shoot that score, so the player would receive an artificially high (less skilled) handicap rating.

The latent variable models presented in this study—particularly the linear mixed effects model from stage one—effectively model player ability and course difficulty simultaneously. This type of model could be used to more accurately and more affordably rate courses and players.[18] This would eliminate biased ratings, would eliminate the need for evaluators to visit courses, and would allow players' handicaps (ability estimates) to update when they have new scores (as they do with the current system). This would require players to enter their scores for every round that they play, which sounds like a potential burden, but this is already a requirement for the handicap rating system, so the new method would not add any new burden onto recreational players.

Dynamic Measurement Modeling

Provides a Tutorial for Conducting DMM Models in R

Most importantly, the tutorials in Chapter Four and Appendix B have demonstrated that DMM models can be estimated in *R*. Previously, all DMM models have been estimated using *SAS*. However, there are significant barriers to using *SAS*, as it is quite expensive. For scholars and practitioners who do not have institutional access to *SAS*, it is likely not a realistic option. Thus, if *SAS* were the only way to estimate a DMM model, many scholars and practitioners would be unable to use these models. This study not only demonstrates that DMM models can be run in *R*, but it provides example code

[18] An example of these difficulty ratings can be found in Appendix C, which shows the 10 easiest courses and 10 most difficult courses according to the model.

on how to do so, including explanations of what the code is doing and what some of the optional settings are.

Because *R* is free and open source, anyone can download it and use it. Providing a tutorial on how to conduct DMM models in *R* should lower (though not eliminate) the barriers to conducting such models for others. This should help increase access to these models and may help them to become more widespread.

Applicability beyond the Field of Education

This study has also demonstrated the applicability of DMM models beyond the field of educational measurement. DMM models were created and originally designed with educational measurement in mind. Thus, this field has dominated the applications of, discussions about, and innovations to DMM in the literature. This makes sense, as the idea of measuring students' capacities in a given domain rather than their current abilities in that domain is naturally appealing.

Nonetheless, there are plenty of other scenarios and other fields of research to which such models may also be useful. This study demonstrates that DMM models can also be successfully applied to sports statistics in order to forecast latent player abilities in the future and to estimate their individual-level latent capacity scores. The same logic could naturally be able to be applied to other sports as well. Even beyond sports, this study has demonstrated that DMM's applicability is not limited to the field of education. This may help it spread to other fields outside of education and even outside of sports statistics. Such spread would increase the prominence and acceptance of this modeling technique. To the extent that DMM models provide benefits over other preexisting

measurement models for some goals, more widespread knowledge and adoption of this technique is desirable.

Extends the Types of DMM Growth Trajectories

Most previous DMM models have been conducted using models with an upper asymptote to represent the player's capacity. In this study, a quadratic model provided a better fit than the more traditional S-shaped and J-shaped models with upper asymptote parameters. Thus, I used a vertex form parameterization of a quadratic growth model to estimate each player's capacity and the time at which the player is estimated to achieve this capacity. This implementation shows that quadratic growth models specifically can be parameterized and conceptualized as having a capacity parameter; they can therefore readily be used as DMM models. More generally, this demonstrates that DMM models should not be limited to the J-shaped and S-shaped growth trajectories that have been used in the past. Any function that has a finite maximum value (likely occurring at a single time point rather than repeatedly) in which the maximum value can be estimated directly as a parameter in the function should be able to be used for the purposes of DMM. Furthermore, in cases when researchers or practitioners are agnostic about the shape of the growth trajectory and simply want to generate reliable and useful capacity scores, this study has demonstrated that they may want to consider a quadratic model as one of their candidate growth trajectories to be tested for best fit.

Continued Demonstration of Benefits of DMM

By showing that the capacity estimates from the DMM model are better at forecasting a player's future maximum ability level than are other single-timepoint

estimates of ability, this study has added to the growing literature on the benefits of DMM as a modeling technique. The DMM model's improvement over single-timepoint observed variable measures was large, while its improvement over single-timepoint latent measures of ability was much smaller. However, even if the improvement was small, the DMM model did still improve upon the latent single-timepoint measure of ability, providing further evidence for previous findings that DMM capacity scores are better able to measure/predict future maximum abilities than are single-timepoint measures.

Future Directions

Based on the models and results presented in this study, there are several directions in which research could proceed from this point. Some of these are methodological while others are substantive. In each case, the new direction would build upon or clarify the results in this study.

Assess the Dimensionality of Golfing Ability

The methods and results that I have presented assume that golfing ability is a unidimensional construct. However, this assumption remains untested in any formal sense, and it is entirely plausible that golfing ability has multiple dimensions (driving ability, putting ability, etc.). Thus, it would be statistically and conceptually valuable to assess the dimensionality of golfing ability using established methods. Such studies would likely use some form of exploratory factor analysis, exploratory item response theory, or structural equation modeling.

If golfing ability does have multiple dimensions, then golf courses should have multiple characteristics as well. Certain courses may favor players who putt better, while

other courses may favor players who have greater driving accuracy, etc. These could then be assessed using two-stage mixed effects models like the process I used in this study or by using structural equation modeling. Modeling the dimensional structure of golfing ability and the differing characteristics of individual courses would naturally be of substantive interest to golf researchers. It would also likely provide a better fit for quantitative models, improving the forecasting ability of models like the ones that I presented.

From a DMM perspective, if golfing ability is multidimensional, then a player could have multiple different capacities that could be estimated using DMM models. These also might occur at different ages. It seems likely that, as they age, players would likely lose their driving distance before their putting ability, and we would be able to model this difference if we had longitudinal data on these different constructs.

Assess Course-Level Discriminations

Even for the unidimensional model presented in this study, it may be useful to assess the discriminations of each course in the dataset rather than assume that they are equally informative. This would require a more complex mixed effects model that would take longer to run, but it would be of both substantive and statistical interest. From a substantive perspective, it would be interesting to know which courses are most effective and least effective at measuring golfing ability, and we could then test whether players seem to (consciously or unconsciously) know this by looking at which tournaments higher-ranked players choose to enter versus those that lower-ranked players choose to enter. Statistically, if some courses are more effective at measuring ability than others

are, modeling this would further improve the accuracy and precision of the ability estimates produced by the model.

Compare Predictive Ability to Multidimensional Measures

In this study, I showed that the unidimensional latent measures of golfing ability produced by the models in the study outperform existing unidimensional measures of golfing ability. Because this included total Strokes Gained, it was already a high bar to clear. It did clear those metaphorical bars, but there would be an even harder test in the form of existing multidimensional measures of golfing ability. In particular, total Strokes Gained is often divided into different dimensions, such as Strokes Gained off the tee, Strokes Gained approaching the green, Strokes Gained putting, etc. If the unidimensional estimates from the models in this study could still outperform the (potentially) more nuanced multidimensional Strokes Gained in terms of prediction, this would yield even further credibility to the latent variable models that I have presented.

Further Analysis to Explain Surprising Tiger Woods Result

The DMM model showed that Tiger Woods, widely regarded as one of the two best players of all time, had only the 11th highest capacity score in the dataset. This is a surprising result. Surprising results themselves are not a problem; indeed, part of the reason for quantitative analyses is to uncover and/or to explain unexpected relationships. However, they do often warrant further examination and explanation. It might be an entirely valid statistical conclusion that Tiger Woods' maximum ability was actually not as high as that of some of the other players. This may imply that the media and fans over-hyped his achievements and/or that he played in an era when the competition was weak,

allowing him to dominate. However, a second option would be that this result is due to some quirk of the data or the model, and that model re-specification or further testing would change the estimate. One possibility in this vein is that scores have generally gotten lower over time due to technology or some factor other than player improvement. If true, the year- and day-varying course difficulty ratings should still capture this. Nonetheless, it would be worth testing. A third option would be that this result is simply due to sampling variability: that the result is a valid conclusion for this particular sample of data but a different (theoretical) sample of data would provide a different conclusion.

Further testing to understand this surprising result would be useful for its own sake and, more importantly, to further confirm the validity of the model. However, the DMM model has already shown its validity. The DMM model did show high reliability during validation, and the correlation between different subsamples and the full sample was high, indicating that option number three is not particularly likely. Similarly, the ability for the DMM model to effectively predict players' maximum ability levels from the stage one model implies that its capacity estimates are valid. Thus, the most likely reason for this surprising result is simply that humans did not understand the world as well as they thought. Nonetheless, further testing could confirm this.

Try Different Growth Trajectories

Now that the quadratic DMM model has been established as a legitimate DMM growth trajectory, further variations of this type of model could be tried. They could be tried with this particular dataset to see if they fit better than the quadratic model, or they could be applied in future DMM studies. For example, one could try an asymmetric

piecewise quadratic function knotted at the vertex, an absolute value function, a piecewise function with an S-shaped growth curve knotted to a quadratic curve at the vertex of the quadratic curve, etc. Buit-in functions in *R* might struggle to find the derivatives of the piecewise options, but these derivatives could be found by hand. The idea of the asymmetric piecewise quadratic function knotted at the vertex is particularly appealing, as it would allow the player's ability levels to increase and decrease at different rates on the two sides of the vertex. Additionally, two of the three (vertex form) parameters would be the same for the two sides of the piecewise function, so it would only add one new parameter to be estimated. This would be similar to the logic of the model estimated by McNeish et al. (2022); they used a quadratic model on the left side of the vertex knotted to a linear function on the right side of the vertex.

Create an *R* Package to Estimate DMM Models

Finally, a major direction forward would be to create an *R* package to estimate DMM models. In this study, I provided a tutorial for estimating DMM models in *R*. This is a significant contribution, but it relies on existing packages that were not originally intended specifically for DMM models. It also still requires the user to be able to write the functional form of the model into R code and figure out reasonable starting values for the estimation. A new *R* package that only required the user to specify the shape of growth curve (exponential, Weibull, quadratic, etc.) from a fixed menu of options rather than coding the equation with all parameters themselves would further lower the barriers to estimating DMM models. This would allow for further implementation by users who want the substantive, statistical, and consequential benefits of DMM models but are not

as comfortable with advanced statistical software or writing code. This, in turn, would assist with the further spread of DMM models.

Concluding Remarks

This study has demonstrated the efficacy of using linear and nonlinear mixed effects models as latent variable measurement models to measure quantities of interest among professional golfers. In particular, this study has measured the latent abilities of professional golfers in specific years using a crossed linear mixed effects model. The player-year ability estimates from this model were then used to estimate a nonlinear Dynamic Measurement Model (DMM) to estimate player capacities: the maximum ability that a player is predicted to have during his career. The results from both stages were successfully validated—the latent measurements consistently outperformed existing observed variable measures of player ability, including Strokes Gained, both in terms of predicting scores at a single timepoint and in terms of forecasting future maximum ability. Furthermore, the resulting estimates showed strong reliability by being consistent across subsamples of the dataset.

These results are noteworthy. By outperforming even the best, most commonly used, and heavily cited existing measures of golfing ability, this study has exceeded a high bar for predictive accuracy. Those wanting to measure golfing ability in the future for its own sake or to be used as independent or dependent variables in future studies would be advised to consider the methods and results detailed in this study. Further refinement in the form of multidimensionality and/or relaxed assumptions can only improve on these estimates.

In addition to the substantive results for golf measurement, these results contribute to the growing methodological literature on Dynamic Measurement Modeling, a relatively new family of latent variable measurement models. Along the way, the study also utilized a new DMM growth trajectory and presented a brief tutorial for future DMM users on how to conduct DMM models using *R*. This provides a new alternative to *SAS* for estimating these models, potentially expanding the universe of potential DMM users and helping these methods gain more widespread recognition and adoption.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22(1), 47–76. https://doi.org/10.3102/10769986022001047

Alexander, D. L., & Kern, W. (2005). Drive for show and putt for dough? An analysis of the earnings of PGA Tour golfers. *Journal of Sports Economics*, 6(1), 46–60. https://doi.org/10.1177/1527002503260797

Aparicio, J., Fried, H., Pastor, J., & Tauer, L. W. (2021). Learning to win on the PGA tour. *Applied Economics*, 53(53), 6104–6119. https://doi.org/10.1080/00036846.2020.1784391

Baker, J., Horton, S., Pearce, W., & Deakin, J. (2006). A longitudinal examination of performance decline in champion golfers. *High Ability Studies*, 16(2), 179–185. https://doi.org/10.1080/13598130600617928

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, 28(2), 135–167. https://doi.org/10.3102/10769986028002135

Baugher, C. D., Day, J. P., & Burford Jr, E. W. (2016). Drive for show and putt for dough? Not anymore. *Journal of Sports Economics*, 17(2), 207–215. https://doi.org/10.1177/1527002514528517

Baumer, B., & Udwin, D. (2015). R Markdown. *WIREs Computational Statistics*, 7(3), 167–177. https://doi.org/10.1002/wics.1348

Belkin, D. S., Gansneder, B., Pickens, M., Rotella, R. J., & Striegel, D. (1994). Predictability and stability of Professional Golf Association tour statistics. *Perceptual and Motor Skills*, 78(3, Suppl.), 1275–1280. https://doi.org/10.2466/pms.1994.78.3c.1275

Berry, S. M. (2001). A statistician reads the sports pages: How ferocious is Tiger? *Chance*, 14(3), 51–56. https://doi.org/10.1080/09332480.2001.10542285

Bliss, A. (2021). Modelling elite golf performance: Predictors of hole score on the European Tour from 2017-2019. *International Journal of Golf Science*, 9(1).

Botha, F., Fraser, G., & Rhoads, T. A. (2021). Skill and earnings amongst golfers on the Southern-African Sunshine Tour. *South African Journal of Economics*, 89(2), 274–281. https://doi.org/10.1111/saje.12269

Bouvet, P. (2011). And if Freddie had been... a new study of the influence of driving and putting on PGA Tour performances. *International Journal of Performance Analysis in Sport*, 11(1), 105–120. https://doi.org/10.1080/24748668.2011.11868533

Broadie, M. (2008). Assessing golfer performance using golfmetrics. *Science and Golf V: Proceedings of the 2008 World Scientific Congress of Golf*, 253–262.

Broadie, M. (2012). Assessing golfer performance on the PGA TOUR. *Interfaces*, 42(2), 146–165. https://doi.org/10.1287/inte.1120.0626

Broadie, M., & Ko, S. (2009). A simulation model to analyze the impact of distance and direction on golf scores. *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 3109–3120. https://doi.org/10.1109/WSC.2009.5429280

Broadie, M., & Rendleman Jr, R. J. (2013). Are the official world golf rankings biased? *Journal of Quantitative Analysis in Sports*, 9(2), 127–140. https://doi.org/10.1515/jqas-2012-0013

Brown, J. (2011). Quitters never win: The (adverse) incentive effects of competing with superstars. *Journal of Political Economy*, 119(5), 982–1013. https://doi.org/10.1086/663306

Callan, S. J., & Thomas, J. M. (2007). Modeling the determinants of a professional golfer's tournament earnings: A multiequation approach. *Journal of Sports Economics*, 8(4), 394–411. https://doi.org/10.1177/1527002506287697

Carsey, T. M., & Harden, J. J. (2013). *Monte Carlo simulation and resampling methods for social science*. Sage Publications.

Chae, J. S., Park, J., & So, W.-Y. (2021). Victory prediction of Ladies Professional Golf Association players: Influential factors and comparison of prediction models. *Journal of Human Kinetics*, 77(1), 245–259.

Chimka, J. R., & Talafuse, T. P. (2016). Poisson regression analysis of additional strokes assessed at golf. *International Journal of Sports Science & Coaching*, 11(4), 619–622. https://doi.org/10.1177/1747954116654785

Clark III, R. D. (2006). The beauty of match play. *Perceptual and Motor Skills*, 102(3), 815–818. https://doi.org/10.2466/pms.102.3.815-818

Clark III, R. D., Woodward, K. L., & Wood, J. M. (2008). On the unreliability of golf scores for professional golfers: A case for restriction of range. *Perceptual and Motor Skills*, 107(3), 683–690. https://doi.org/10.2466/pms.107.3.683-690

Clarke, S. R., Rice, J. M., & others. (1995). How well do golf courses measure golf ability? An application of test reliability procedures to golf tournament scores. *ASOR BULLETIN*, 14, 2–11.

Coate, D., & Toomey, M. (2014). Do professional golf tour caddies improve player scoring? *Journal of Sports Economics*, 15(3), 303–312. https://doi.org/10.1177/1527002512458799

Connolly, R. A., & Rendleman Jr, R. J. (2008). Skill, luck, and streaky play on the PGA tour. *Journal of the American Statistical Association*, 103(481), 74–88. https://doi.org/10.1198/016214507000000310

Connolly, R. A., & Rendleman Jr, R. J. (2011). Going for the green: A simulation study of qualifying success probabilities in professional golf. *Journal of Quantitative Analysis in Sports*, 7(4). https://doi.org/10.2202/1559-0410.1308

Connolly, R. A., & Rendleman Jr, R. J. (2012). What it takes to win on the PGA TOUR (if your name is "Tiger" or if it isn't). *Interfaces*, 42(6), 554–576. https://doi.org/10.1287/inte.1110.0615

Connolly, R., & Rendleman Jr, R. J. (2012). Tournament selection efficiency: An analysis of the PGA TOUR's FedExCup. *Journal of Quantitative Analysis in Sports*, 8(4). https://doi.org/10.1515/1559-0410.1495

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38(4), 529–569. https://doi.org/10.1207/s15327906mbr3804_5

Davidson, J. D., & Templin, T. J. (1986). Determinants of success among professional golfers. *Research Quarterly for Exercise and Sport*, 57(1), 60–67. https://doi.org/10.1080/02701367.1986.10605389

Dong, Y., & Dumas, D. (2024). *Validation practices for dynamic measurement*.

Dong, Y., Dumas, D., Clements, D. H., & Sarama, J. (2022). Developing a trajectory deviance index for dynamic measurement modeling. *The Journal of Experimental Education*, 1-22. https://doi.org/10.1080/00220973.2022.2044280

Dorsel, T. N., & Rotunda, R. J. (2001). Low scores, top 10 finishes, and big money: An analysis of professional golf association tour statistics and how these relate to overall performance. *Perceptual and Motor Skills*, 92(2), 575–585. https://doi.org/10.2466/pms.2001.92.2.575

Drappi, C., & Co Ting Keh, L. (2019). Predicting golf scores at the shot level. *Journal of Sports Analytics*, 5(2), 65–73. https://doi.org/10.3233/JSA-170273

Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: Using nonlinear growth models to estimate student learning capacity. *Educational Researcher*, 46(6), 284-292. https://doi.org/10.3102/0013189X17725747

Dumas, D. G., & McNeish, D. M. (2018). Increasing the consequential validity of reading assessment using dynamic measurement modeling: A comment on Dumas and McNeish (2017). *Educational Researcher*, 47(9), 612-614. https://doi.org/10.3102/0013189X18797621

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical–psychometric paradigm for modern educational psychology. *Educational Psychologist*, 55(2), 88-105. https://doi.org/10.1080/00461520.2020.1744150

Dumas, D., McNeish, D., Sarama, J., & Clements, D. (2019). Preschool mathematics intervention can significantly improve student learning trajectories through elementary school. *AERA Open*, 5(4), 1-15. https://doi.org/10.1177/2332858419879446

Dumas, D., McNeish, D., Schreiber-Gregory, D., During, S. J., & Torre, D. M. (2019). Dynamic measurement in health professions education: Rationale, application, and possibilities. *Academic Medicine*, 94(9), 1323-1328. https://doi.org/10.1097/ACM.0000000000002729

Elmore, R., & Urbaczewski, A. (2018). Hot and cold hands on the PGA Tour: Do they exist? *Journal of Sports Analytics*, 4(4), 275–284. https://doi.org/10.3233/JSA-180214

Engelhardt, G. M. (1995). “It’s not how you drive, it’s how you arrive”: The myth. *Perceptual and Motor Skills*, 80(3, Suppl.), 1135–1138. https://doi.org/10.2466/pms.1995.80.3c.1135

Engelhardt, G. M. (1997). Differences in shot-making skills among high and low money winners on the PGA tour. *Perceptual and Motor Skills*, 84(3, Suppl.), 1314. https://doi.org/10.2466/pms.1997.84.3c.1314

Engelhardt, G. M. (1999). Is American PGA tour more competitive than European tour? Reply to Jiménez and Fierro-Hernández. *Perceptual and Motor Skills*, 89(3), 1028. https://doi.org/10.2466/pms.1999.89.3.1028

Engelhardt, G. M. (2002). Driving distance and driving accuracy equals total driving: Reply to Dorsel and Rotunda. *Perceptual and Motor Skills*, 95(2), 423–424. https://doi.org/10.2466/pms.2002.95.2.423

Fearing, D., Acimovic, J., & Graves, S. C. (2011). How to catch a Tiger: Understanding putting performance on the PGA Tour. *Journal of Quantitative Analysis in Sports*, 7(1). https://doi.org/10.2202/1559-0410.1268

Feuerstein, R., Feuerstein, R., & Falik, L. H. (2015). *Beyond smarter: Mediated learning and the brain's capacity for change*. Teachers College Press.

Feuerstein, R., Rand, Y., & Hoffman, M. (1979). *The dynamic assessment of retarded performers: The learning potential, assessment device, theory, instruments and techniques*. University Park Press.

Finley, P. S., & Halsey, J. J. (2004). Determinants of PGA tour success: An examination of relationships among performance, scoring, and earnings. *Perceptual and Motor Skills*, 98(3), 1100–1106. https://doi.org/10.2466/pms.98.3.1100-1106

Fisher, P. (1998). Objective analysis of golf. *Popular Measurement*, 1(1), 41–42.

Fried, H. O., Lambrinos, J., & Tyner, J. (2004). Evaluating the performance of professional golfers on the PGA, LPGA and SPGA tours. *European Journal of Operational Research*, 154(2), 548–561. https://doi.org/10.1016/S0377-2217(03)00188-7

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications.

Hamel, S., Caudill, S. B., & Mixon Jr, F. G. (2016). A good walk foiled: Monopoly power and barriers to entry into the PGA Tour. *Managerial and Decision Economics*, 37(8), 574–584. https://doi.org/10.1002/mde.2752

Heiny, E. L. (2008). Today's PGA Tour pro: Long but not so straight. *Chance*, 21(1), 10–21. https://doi.org/10.1080/09332480.2008.10722880

Heiny, E. L., & Frisby, C. C. (2018). An ordinal logistic regression model for the Masters Golf Tournament. *Chance*, 31(3), 44–58. https://doi.org/10.1080/09332480.2018.1522213

Heiny, E. L., & Heiny, R. (2012a). And the 2011 driving champion is? Dustin Johnson. *Journal of Quantitative Analysis in Sports*, 8(4). https://doi.org/10.1515/1559-0410.1476

Heiny, E. L., & Heiny, R. L. (2012b). An 11-year study of the relative importance of driving accuracy and driving distance on scoring average on the PGA tour. *Chance*, 25(1), 4–14. https://doi.org/10.1080/09332480.2012.668457

Heiny, E. L., & Heiny, R. L. (2014). Stochastic model of the 2012 PGA Tour season. *Journal of Quantitative Analysis in Sports*, 10(4), 367–379. https://doi.org/10.1515/jqas-2014-0043

Hoegh, A. (2011). Defining the performance coefficient in golf: A case study at the 2009 Masters. *Journal of Quantitative Analysis in Sports*, 7(2). https://doi.org/10.2202/1559-0410.1331

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

James, N. (2007). The statistical analysis of golf performance. *International Journal of Sports Science & Coaching*, 2(1, Suppl.), 231–249. https://doi.org/10.1260/174795407789705424

James, N. (2009). Performance analysis of golf: Reflections on the past and a vision of the future. *International Journal of Performance Analysis in Sport*, 9(2), 188–209. https://doi.org/10.1080/24748668.2009.11868476

James, N., & Rees, G. D. (2008). Approach shot accuracy as a performance indicator for US PGA Tour golf professionals. *International Journal of Sports Science & Coaching*, 3(1, Suppl.), 145–160. https://doi.org/10.1260/174795408785024225

Jiménez, J. A., & Fierro-Hernández, C. (1999). Are European and American golf players different? Reply to Engelhardt (1997). *Perceptual and Motor Skills*, 89(2), 417–418. https://doi.org/10.2466/pms.1999.89.2.417

Jonsson, E. N., Karlsson, M. O., & Wade, J. R. (2000). Nonlinearity detection: Advantages of nonlinear mixed-effects modeling. *AAPS PharmSci*, 2(3), 114–123. https://doi.org/10.1208/ps020332

Kahane, L. H. (2010). Returns to skill in professional golf: A quantile regression approach. *International Journal of Sport Finance*, 5(3), 167.

Kamata, A. (2001). Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement*, 38(1), 79–93. https://doi.org/10.1111/j.1745-3984.2001.tb01117.x

Karlsen, J., & Nilsson, J. (2008). Distance variability in golf putting among highly skilled players: The role of green reading. *International Journal of Sports Science & Coaching*, 3(1, Suppl.), 71–80. https://doi.org/10.1260/174795408785024333

Ketzcher, R., & Ringrose, T. J. (2002). Exploratory analysis of European professional golf association statistics. *Journal of the Royal Statistical Society Series D: The Statistician*, 51(2), 215–228. https://doi.org/10.1111/1467-9884.00313

Kim, E.-K., & Chae, J.-S. (2021). Comparison of the performance levels of US female professional golf players. *International Journal of Applied Sports Sciences*, 33(1). https://doi.org/10.24985/ijass.2021.33.1.61

Kirschenbaum, R. J. (1998). Dynamic assessment and its use with underserved gifted and talented populations. *Gifted Child Quarterly*, 42(3), 140-147. https://doi.org/10.1177/001698629804200302

Korpimies, S. (2020). *Predicting players' success on the PGA-Tour* [Master's Thesis]. Aalto University.

Kuhn, E., & Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4), 1020–1038. https://doi.org/10.1016/j.csda.2004.07.002

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Lamb, P., Stöckl, M., & Lames, M. (2011). Performance analysis in golf using the ISOPAR method. *International Journal of Performance Analysis in Sport*, 11(1), 184–196. https://doi.org/10.1080/24748668.2011.11868539

Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research*, 15(1), 11-33. https://doi.org/10.1177/1362168810383328

Leahy, B. (2014). *Predicting professional golfer performance using proprietary PGA Tour "Shotlink" data* [Master's Thesis]. Technological University Berlin.

Lee, S. Y., Lei, B., & Mallick, B. (2020). Estimation of COVID-19 spread curves integrating global data and borrowing information. *PLOS ONE*, 15(7), e0236860. https://doi.org/10.1371/journal.pone.0236860

Lindstrom, M. J., & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3), 673. https://doi.org/10.2307/2532087

Manwaring, T. C. (2016). *A comparison in the returns to skills on the European Tour and the PGA Tour* [Senior Thesis]. Skidmore College.

McLaughlin, K., & Cascella, P. W. (2008). Eliciting a distal gesture via dynamic assessment among students with moderate to severe intellectual disability. *Communication Disorders Quarterly*, 29(2), 75-81. https://doi.org/10.1177/1525740107311821

McNeish, D., & Dumas, D. (2017). Nonlinear growth models as measurement models: A second-order growth curve model for measuring potential. *Multivariate Behavioral Research*, 52(1), 61-85. https://doi.org/10.1080/00273171.2016.1253451

McNeish, D., & Dumas, D. (2018). Calculating conditional reliability for dynamic measurement model capacity estimates. *Journal of Educational Measurement*, 55(4), 614-634. https://doi.org/10.1111/jedm.12195

McNeish, D., & Dumas, D. (2021). A seasonal dynamic measurement model for summer learning loss. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2), 616-642. https://doi.org/10.1111/rssa.12634

McNeish, D., & Dumas, D. G. (2019). Scoring repeated standardized tests to estimate capacity, not just current ability. *Policy Insights from the Behavioral and Brain Sciences*, 6(2), 218-224. https://doi.org/10.1177/2372732219862578

McNeish, D., Dumas, D. G., & Grimm, K. J. (2020). Estimating new quantities from longitudinal test scores to improve forecasts of future performance. *Multivariate Behavioral Research*, 55(6), 894-909. https://doi.org/10.1080/00273171.2019.1691484

McNeish, D., Dumas, D., Torre, D., & Rice, N. (2022). Modelling time to maximum competency in medical student progress tests. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(4), 2007–2034. https://doi.org/10.1111/rssa.12864

Moy, R. L., & Liaw, T. (1998). Determinants of professional golf tournament earnings. *The American Economist*, 42(1), 65–70. https://doi.org/10.1177/056943459804200106

M.R. Farrally A.J. Cochran, D.J. Crews, M.J. Hurdzan, R.J. Price, J.T. Snow, & P.R. Thomas (2003). Golf science research at the beginning of the twenty-first century. *Journal of Sports Sciences*, 21(9), 753–765. https://doi.org/10.1080/0264041031000102123

Nix, C. L., & Koslow, R. (1991). Physical skill factors contributing to success on the professional golf tour. *Perceptual and Motor Skills*, 72(3, Suppl.), 1272–1274. https://doi.org/10.2466/pms.1991.72.3c.1272

Ohn, J. K., Bealing, W., & Waeger, D. (2012). The determinants of annual earnings for PGA Players under the new PGA's FedEx Cup system. *Review of Applied Economics*, 8(1), 95–105. http://dx.doi.org/10.22004/ag.econ.143466

Park, I., & Lee, Y. H. (2012). Efficiency comparison of international golfers in the LPGA. *Journal of Sports Economics*, 13(4), 378–392. https://doi.org/10.1177/1527002512450263

Pinheiro, J., Bates, D., & R Core Team. (2022). *nlme: Linear and nonlinear mixed Effects models*. https://cran.r-project.org/package=nlme

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer. https://doi.org/10.1007/b98882

Pope, D. G., & Schweitzer, M. E. (2011). Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes. *American Economic Review*, 101(1), 129-157. https://doi.org/10.1257/aer.101.1.129

Preacher, K. J., & Hancock, G. R. (2015). Meaningful aspects of change as novel random coefficients: A general method for reparameterizing longitudinal models. *Psychological Methods*, 20(1), 84. https://doi.org/10.1037/met0000028

R Core Team. (2022). *R: A language and environment for statistical computing*. https://www.r-project.org/

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, 69(2), 167–190. https://doi.org/10.1007/BF02295939

Raket, L. L. (2020). Statistical disease progression modeling in Alzheimer Disease. *Frontiers in Big Data*, 3, 1–18. https://doi.org/10.3389/fdata.2020.00024

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Second Edition, Vol. 1). Sage.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8(2), 185–205. https://doi.org/10.1037/1082-989X.8.2.185

Rinehart, K. L. (2009). The economics of golf: An investigation of the returns to skill of PGA tour golfers. *Major Themes in Economics*, 11(1), 57–70.

Ritz, C., Baty, F., Streibig, J. C., & Gerhard, D. (2015). Dose-response analysis using R. *PLOS ONE*, 10(12), e0146021. https://doi.org/10.1371/journal.pone.0146021

Sachau, D., Andrews, L., Gibson, B., & DeNeui, D. (2009). Tournament validity: Testing golfer competence. *Measurement in Physical Education and Exercise Science*, 13(1), 52–69. https://doi.org/10.1080/10913670802611017

SAS Software. (n.d.). SAS Institute.

Sen, K. C. (2012). Mapping statistics to success on the PGA Tour: Insights from the use of a single metric. *Sport, Business and Management: An International Journal*, 2(1), 39–50. https://doi.org/10.1108/20426781211207656

Sharma, A., & Reilly, P. (2013). A comparative study of the indicators of success on the PGA Tour: A panel data analysis. *International Journal of Economic Practices and Theories*, 3(1), 29–36.

Shmanske, S. (1992). Human capital formation in professional sports: Evidence from the PGA Tour. *Atlantic Economic Journal*, 20, 66–80. https://doi.org/10.1007/BF02300173

Shmanske, S. (2008). Skills, performance, and earnings in the tournament compensation model: Evidence from PGA Tour microdata. *Journal of Sports Economics*, 9(6), 644–662. https://doi.org/10.1177/1527002508317469

Shmanske, S. (2009). Golf match: the choice by PGA Tour golfers of which tournaments to enter. *International Journal of Sport Finance*, 4(2), 114–135.

Stemen, G. A. (2002). *Distance off the tee and its impact on scoring average compared to accuracy off the tee and its impact on scoring average* [Master's Action Research Project]. Southwest State University.

Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C., Jukes, M., & Bundy, D. A. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence*, 30, 141-162. https://doi.org/10.1016/S0160-2896(01)00091-5

Stigler, S. M., & Stigler, M. L. (2018). Luck and skill in tournament golf. *CHANCE*, 31(3), 4–13. https://doi.org/10.1080/09332480.2018.1522206

Stöckl, M., Lamb, P. F., & Lames, M. (2011). The ISOPAR method: A new approach to performance analysis in golf. *Journal of Quantitative Analysis in Sports*, 7(1). https://doi.org/10.2202/1559-0410.1289

Stöckl, M., Lamb, P. F., & Lames, M. (2012). A model for visualizing difficulty in golf and subsequent performance rankings on the PGA Tour. *International Journal of Golf Science*, 1(1), 10–24.

Team, R. C. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Tjørve, K. M. C., & Tjørve, E. (2017). The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the unified-Richards Family. *PloS One*, 12(6), 1–17. https://doi.org/10.1371/journal.pone.0178691

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2), 225–255. https://doi.org/10.1348/000711005X79857

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28(4), 369–386. https://doi.org/10.3102/10769986028004369

Watkins, J. R. , Jr. (2008). Drive for show, putt for dough: Rates of return to golf skills, events played, and age on the PGA Tour. *Michigan Journal of Business*, 1, 35–59.

Wiseman, F., & Chatterjee, S. (2006). Comprehensive analysis of golf performance on the PGA Tour: 1990–2004. *Perceptual and Motor Skills*, 102(1), 109–117. https://doi.org/10.2466/pms.102.1.109-117

Appendix A: Golf Terminology

Golf is a sport played by trying to hit a ball on the ground with a stick (or 'club') into a small hole in the ground, which is typically hundreds of yards away. The golfer tries to minimize the number of times he or she hits the ball (called 'shots' or 'strokes'). This process is then repeated 18 times to complete one round of golf. The goal is to minimize the number of shots that it takes to complete all 18 holes.

The first shot on each hole is hit from the tee box, a patch of grass that is typically cut shorter than the surrounding areas. It has two tee markers, and the golfer can select where between these markers to place the ball before hitting the shot. After this shot, the ball lands on the green, fairway, rough, or sand trap. Wherever it lands, the player has to play the next shot from the position in which the ball comes to rest.

Each golf hole is given a par value, which is, in theory, the expected number of shots that it would take a 'competent' golfer to get the ball into the hole. In practice, however, the par value is based mostly on the length of the hole rather than its true difficulty. Par 3 holes are short enough that a single good shot should be able to put the ball on the green. A par 4 will generally require two good shots to get the ball on the green. A par 5 will generally require three good shots or two excellent ones to get onto the green.

Once on the green (which has the shortest grass on the course), a player is putting. A putt is a shot in which the ball leaves the clubface immediately rolling along the ground without flying through the air first. If the player is near the green, but not in it, he

or she will be chipping or pitching (shots that do not require full swings) to get the ball near the hole. Once a hole is complete, the player moves on to the next hole.

The following terms may be useful for understanding the sport of golf:

Approach Shot: See approaching the green.

Approaching the Green: An approach shot is a shot from the fairway, rough, or fairway bunker on a par 4 or par 5 or from the tee box on a par 3 that is intended to get the ball onto the green. Approach shot ability refers to the skill that a player is able to utilize to accomplish this task.

Ball: the golf ball used under the rules of golf is restricted in several ways. It can be at most 1.62 ounces. It must be spherical (although small indentations called “dimples” are allowed), and it must be symmetric. Golf balls are traditionally white, though other colors are allowed and used occasionally.

Bounce back: A bounce back occurs when a player immediately follows an over-par hole with an under-par hole.

Bunker: See sand trap

Chipping: A chip is a shot from near the green that is intended to get the ball near the hole by allowing the ball to cover a significant percentage of the distance on the ground rather than through the air. Correspondingly, these shots stay low to the ground, and the golfer typically does not break his or her wrists. This is in contrast to pitching.

Club: The object used by the player to strike the ball. Players utilize clubs of different lengths and lofts depending on the needs of the particular shot. The maximum number of clubs that a player can carry in a particular round is 14.

Cup: See hole

Cut: "cut" can mean two different things in the context of golf. First, it can refer to the scenario in which a player hits the ball with sidespin that causes the ball to spin left to right (for a right-hander) or right to left (for a left-hander). The second way that "cut" is used refers to the winnowing of the field in golf tournaments between the second and third round. On the PGA Tour, most tournaments allow the top 65 players and ties to make the cut, while those further behind the leader are cut.

Driving: In the context of golf, driving refers to the tee shot (first shot of the hole) on par 4 and par 5 holes.

Fairway: A portion of the hole in between the tee box and the green. The grass is typically cut shorter than the rough but longer than on the green or on the tee box. The fairway is generally considered a desirable landing spot for a tee shot on par 4 and par 5 holes because it makes the approach shot to the green easier.

Green: The area closest to the hole. The grass is cut the shortest on this part of the course so that the player is able to roll the ball without hitting it through the air ("putting").

Green in Regulation (GIR): A green in regulation refers to a situation in which a player hits the ball onto the green in two less than the par of the hole. Thus, a green in regulation would mean getting the ball onto the green in one shot on a par 3, in two shots (or fewer) on a par 4, and in three shots (or fewer) on a par 5.

Handicap Rating System: A system used to rate the relative skill levels of golfers, mostly used for amateur golfers rather than professional golfers. A player with a handicap rating of 13 would be expected to finish the round 13 strokes above par on a course of average difficulty. This system allows golfers of varying abilities to compete on a theoretically level playing field.

Hole: A golf hole refers to two different concepts. First, the "hole" refers to the physical hole in the ground into which players try to hit the ball. This may also be called the "cup." Second, the "hole" refers to the entire area from tee box to the green. The golf

hole begins by hitting a tee shot from the tee box. The player is allowed to use a tee for this first shot and can choose the spot from which to hit (between the two tee markers). The player then hits subsequent shots until getting the ball in the hole. Most holes are par 3s, par 4s, or par 5s. A complete round of golf consists of 18 holes, with the round score being the sum of the 18 individual hole scores.

LPGA Tour: The Ladies Professional Golf Association. This is the most prestigious and lucrative of the women's professional golf tours. Most, though not all, tournaments are held in the United States.

Money List: A method of ranking players based solely on their prize money earnings in a given season. It was frequently used in the past, but it has mostly been replaced by more formal rankings systems.

Official World Golf Rankings (OWGR): A points-based system that ranks players on multiple tours.

Par: The number of shots in which a proficient golfer would be expected to complete a hole. This is based in large part on the length of the hole. For example, a "Par 3" would be a hole in which a proficient golfer would be expected to take 3 shots to get the ball from tee box into the hole. Almost all golf holes have a par value of 3, 4, or 5.

PGA European Tour: The main professional golf tour based in Europe, it is also called the DP World Tour, among other names. Most, though not all, tournaments take place in Europe.

PGA Tour: The most prestigious of the men's professional golf tours. Most, though not all, tournaments take place in the United States.

PGA Tour Champions: The most prestigious of the men's professional golf tours dedicated to golfers of age 50+. Most, though not all, tournaments take place in the United States.

Pitching: A pitch is a shot from near the green that is intended to get the ball near the hole by carrying a significant percentage of the distance through the air rather than on the ground. Correspondingly, these shots have relatively high loft, and the golfer typically uses more wrist action than when chipping.

Putting: Shots when the ball is on the green

Rough: The area on a golf hole surrounding the fairway. The grass is longer than the fairway, and this is considered a less desirable location from which to approach the green (and therefore a less desirable landing spot for a tee shot) on par 4 and par 5 holes

Round: A round of golf consists of 18 holes, often divided into a "front 9" and "back 9." A player's score for the round is the sum of the 18 individual hole score.

Sand Save: When a ball is in a greenside sand trap, a player is said to have a sand save if the player is able to get the ball into the hole in two shots (or less). Sand save percentage refers to the percentage of the time that a player is able to accomplish this.

Sand Traps: Also known as "bunkers." These are portions of a golf hole that are covered in sand. The ground is typically also somewhat below the level of the surrounding fairway, rough, or green. A Greenside bunker is one next to the green. A fairway bunker is one that is not next to the green.

Scrambling: Scrambling refers to saving par (or better) on a hole when the player does not get the ball onto the green in regulation. Scrambling percentage refers to the percentage of the time that a player is able to achieve this.

Shot/Stroke: When a player swings with the club and attempts to make contact with the ball, this counts as one shot or one stroke.

Sunshine Tour: A developmental golf tour based in South Africa. As a lower-level tour, most players try to "graduate" from this tour to the European Tour.

Tee: In addition to referring to the tee box, a tee is a small wooden or plastic stand on which a golfer is allowed to put the ball on the first shot of each hole.

Tee Box: The start of a hole, usually with grass cut lower than that of the surrounding areas. The tee box has two tee markers, and players get to choose from where between those markers they would like to hit the first shot.

Tee Shot: A shot from the tee box. Similar to driving, but it also includes the first shot on par 3 holes.

Tournament: A golf tournament is played over 3-5 rounds (usually 4). The player with the lowest total score over these rounds wins the tournament. Often, after the second round, the field is cut.

Appendix B: Example Code for other DMM Trajectories

```
library(foreign)
data.dmm<-read.csv("C:/Users/macwe/OneDrive/Dissertation/Datasets
/player-year.csv", sep="")
data.dmm[c(1:10),] #same dataset as in Chapter 4
```

```
##            PLAYER Year      sum1 Age
## 6     A Ilyassyak 2007 -1.884216  42
## 7     A Ilyassyak 2008 -1.211551  43
## 8     A Ilyassyak 2009 -1.320875  44
## 14     A Siddikur 2010  2.453366  26
## 15     A Siddikur 2011  2.552817  27
## 22    A.J. Crouch 2022  2.896804  29
## 23    A.J. Crouch 2023  2.295983  30
## 24    A.J. Elgert 2008  1.837466  26
## 25    A.J. Elgert 2010  2.270091  28
## 27 A.J. McInerney 2017  2.453366  24
```

```
###### Michaelis-Menten Model
library(drc)
mod.mm<-drm(sum1~Age,data=data.dmm,fct=drc::MM.3(fixed=c(NA,NA,NA
), names=c("B0","Bc","Bm")))
summary(mod.mm)
```

```
##
## Model fitted: Shifted Michaelis-Menten (3 parms)
##
## Parameter estimates:
##
##                   Estimate  Std. Error t-value p-value
## B0:(Intercept)  2.4692e+00  3.1998e-02 77.1671  <2e-16 ***
## Bc:(Intercept) -6.5454e+02  5.2998e+02 -1.2350  0.2168
## Bm:(Intercept)  2.4249e+04  1.9615e+04  1.2363  0.2164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error:
##
##  1.711113 (29420 degrees of freedom)
```

```
library(lme4)
nform<- ~0+B0+(((Bc-B0)*input)/(Bm+input))
nfun<-deriv(nform, namevec=c("B0","Bc","Bm"), function.arg=c("inp
ut", "B0", "Bc","Bm"))
c.dmm<-nlmerControl(optimizer="bobyqa",tolPwrss=10^-9,optCtrl=lis
t(rhobeg=0.001,rhoend=10^-8,maxfun=40000))
#model.dmm.mm<-nlmer(sum1~nfun(Age,B0,Bc,Bm)~B0+Bc+Bm+(0+Bc | PLA
YER), data=data.dmm, start = c(B0=2.4692,Bc=-0.065454,Bm=24249),c
ontrol=c.dmm)
# Does not converge

##### Exponential Model
mod.exp<-drm(sum1~Age,data=data.dmm,fct=drc::EXD.3(fixed=c(NA,NA,
NA),names=c("B0","Bc","Br")))
summary(mod.exp)

##
## Model fitted: Shifted exponential decay (3 parms)
##
## Parameter estimates:
##
##                    Estimate  Std. Error t-value p-value
## B0:(Intercept) -3.3925e+02  2.0499e+02 -1.6550 0.09794 .
## Bc:(Intercept)  2.4711e+00  3.1974e-02 77.2850 < 2e-16 ***
## Br:(Intercept)  1.2589e+04  7.5660e+03  1.6639 0.09614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error:
##
##  1.711111 (29420 degrees of freedom)

library(lme4)
nform<- ~0+B0+(Bc-B0)*(1-exp(Br*input))
nfun<-deriv(nform, namevec=c("B0","Bc","Br"), function.arg=c("inp
ut", "B0", "Bc","Br"))
c.dmm<-nlmerControl(optimizer="bobyqa",tolPwrss=10^-9,optCtrl=lis
t(rhobeg=0.001,rhoend=10^-8,maxfun=40000))
#model.dmm.e<-nlmer(sum1~nfun(Age,B0,Bc,Br)~B0+Bc+Br+(0+Bc | PLAY
ER), data=data.dmm, start =c(B0=-339.25,Bc=2.4711,Br=(-1/12589)),
control=c.dmm)
#Does not Converge

#### Logistic
```

```
mod.l<-drm(sum1~Age,data=data.dmm,fct=drc::L.4(fixed=c(NA,NA,NA,NA), names=c("Br","B0","Bc","Bm")))
summary(mod.l)

## Warning in sqrt(diag(varMat)): NaNs produced

##
## Model fitted: Logistic (ED50 as parameter) (4 parms)
##
## Parameter estimates:
##
##                    Estimate  Std. Error t-value   p-value
## Br:(Intercept)  8.5466e-02  2.2797e-03   37.49 < 2.2e-16 ***
## B0:(Intercept) -3.3001e+02         NaN     NaN       NaN
## Bc:(Intercept)  1.8139e+00  1.6022e-02  113.21 < 2.2e-16 ***
## Bm:(Intercept)  1.2426e+02         NaN     NaN       NaN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error:
##
##  1.681344 (29419 degrees of freedom)

################## Did not run DMM version because marginal model already too complex
#########
#library(lme4)
#nform<- ~0+BL+((Bc-BL)/(1+exp(Br*(input-Bm))))
#nfun<-deriv(nform, namevec=c("BL","Bc","Br","Bm"), function.arg=c("input", "BL", "Bc","Br","Bm"))
#c.dmm<-nlmerControl(optimizer="bobyqa",tolPwrss=10^-9,optCtrl=list(rhobeg=0.001,rhoend=10^-8,maxfun=40000))
#model.dmm.l<-nlmer(sum1~nfun(Age,BL,Bc,Br,Bm)~BL+Bc+Br+Bm+(0+Bc | PLAYER), data=data.dmm, start = c(B0=0,Bc=0,Br=0,Bm=0,control=c.dmm)

#### Weibull
mod.w<-drm(sum1~Age,data=data.dmm,fct=drc::W2.4(fixed=c(NA,NA,NA,NA), names=c("Br","B0","Bc","Bi")))
summary(mod.w)

##
## Model fitted: Weibull (type 2) (4 parms)
##
```

```
## Parameter estimates:
##
##                 Estimate Std. Error  t-value   p-value
## Br:(Intercept) -4.929984   0.588907  -8.3714 < 2.2e-16 ***
## B0:(Intercept) -1.612795   0.435576  -3.7027 0.0002137 ***
## Bc:(Intercept)  1.725544   0.011187 154.2503 < 2.2e-16 ***
## Bi:(Intercept) 56.305278   1.686485  33.3862 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error:
##
##  1.672485 (29419 degrees of freedom)
```

```
nform<- ~0+Bc+(BL-Bc)*(1-exp(-exp(Br*(log(input)-log(Bi)))))
nfun<-deriv(nform, namevec=c("BL","Bc","Br","Bi"), function.arg=c
("input", "BL", "Bc","Br","Bi"))
c.dmm<-nlmerControl(optimizer="bobyqa",tolPwrss=10^-9,optCtrl=lis
t(rhobeg=0.001,rhoend=10^-8,maxfun=40000))
#model.dmm.ww<-nlmer(sum1~nfun(Age,BL,Bc,Br,Bi)~BL+Bc+Br+Bi+(0+Bc
| PLAYER), data=data.dmm, start = c(BL=-1.612795,Bc=1.725544,Br=-
4.929984,Bi=56.305278),control=c.dmm)
### Does not Converge
```

Appendix C: Example Course Difficulties

Table 13: Hardest Courses

Course Name	Location	Shots above Average
The Ocean Course at Kiawah Island	South Carolina, USA	3.384
Oakmont Country Club	Pennsylvania, USA	3.316
Augusta National Golf Club	Georgia, USA	3.264
Shinnecock Hills Golf Club	New York, USA	3.217
Winged Foot Golf Club	New York, USA	3.140
Merion Golf Club	Pennsylvania, USA	3.004
Hazeltine National Golf Club	Minnesota, USA	2.772
The Course at Wente Vineyards	California, USA	2.613
Adare Manor Golf Club	County Limerick, Ireland	2.602
Muirfield Village Golf Club	Ohio, USA	2.570

Table 14: Easiest Courses

Course Name	Location	Shots above Average
Royal Durban Golf Club	Durban, South Africa	-4.224
Bogota Country Club (Pacos Course)	Bogota, Colombia	-2.803
Fontainebleau Golf Club	Fontainebleau, France	-2.553
Eisenhower Park (Red Course)	New York, USA	-2.498
Canyon Meadows Golf and Country Club	Alberta, Canada	-2.456
Meloneras Golf Club	Las Palmas, Spain	-2.384
Augusta Pines Golf Club	Texas, USA	-2.371
Champions Run Golf Course	Tennessee, USA	-2.301
Royal Johannesburg (West Course)	Johannesburg, South Africa	-2.261
Oakridge Country Club	Utah, USA	-2.254