ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis [Economics and Business Economics]

The effect of draw and rewards of an ATP tour single-elimination tournament on win chances of the top seed player

Name student: Sebastiaan Tijmensen

Student ID number: 482625

Supervisor: Sisak, D.

Second assessor: Crutzen, B.S.Y.

Date final version: 08/31/2020

Credits:
A big thank you to dr. Dana Sisak for the supervision of this thesis in these difficult times. This would not have been possible without her knowledge, feedback and support.

Abstract

This paper aimed to add variables to the already existing variables in predicting tennis outcomes pre-match. Raw ATP Tour data is used, comparing 3945 matches of top seed players between 1991 and 2016. The data is used to research matches of top seed players, because the top seed is the most important player for a tournament director in the sense of publicity and revenue. First, the already known variables from previous literature were put into a Probit model on the match outcomes of the top seed player. Secondly, (dummy) variables accounting for the impact of byes, specific tournament rounds and draw sizes were implemented into the Probit model. Results support the significance of variables added from previous literature and signal a decrease in win chances, *ceteris paribus,* with draw size increasing or a tournament being more rewarding. Top seed players do in reality win more matches and tournaments if a tournament is more prestigious, but this is found to be, mainly, due to a consequent increase in prize money and/or the presence of byes.

Table of contents

In recent years, ABN AMRO World Tennis Tournament has chosen the strategy of not having a “superstar” player in the playing field but to have as much top 20 players as possible and therefore have more close games at a high level (Rijnmond, 2020). Tournament director Richard Krajicek has to make this decision, as these top tier players ask a higher appearance fee. Having to pay higher fees, would cancel out the opportunity to sign more quality players. Lack of top players leads to playing sessions without an eye-catching player.

Chmait, Robertson, Westerbeek, Eime, Sellitto and Reid (2020) came to the insight that famous players playing, does have a significant positive influence on the sale of tickets during a tennis tournament. In this sense, it is very important for a tournament to have exiting players in all sessions and have the most attractive players reach the latter stages of the tournament to generate the highest possible revenue.

In a tennis tournament, the top seed player is generally the most attractive player in the field. The top seed is expected to have the greatest skills, as seeding is based on the world rankings, and the rankings are based on past results. A tournament director would want this player to play as many matches as possible. If there are exploitable variables in favor of the top seed player, this could be highly interesting for the directors of tennis tournaments. The betting industry, firm CEO’s, and other sports could also have interests in influenceable variables that create favorable circumstances for certain competitors.

Employers always try to give incentives to employees to encourage optimum productivity. Lazear and Rosen (1979) came with the idea of rewarding employees on their rank in the company. People compete in effectiveness during their jobs to achieve promotion to a better paid job. They found effort to be increasing if the spread of money between winning and losing became larger and found effort decreasing if the factor of luck became more influential. DeVaro (2006) adds that promotion is based on relative performance (beating opponents). Employers want the most effective employees near the end of the tournament (more important jobs) just like tennis tournaments want the best players to compete in the end stages. Wage spread and luck involved could be incentives for the favorable player to exert a different level of effort in tennis tournaments and in intrafirm tournaments.

Variables predicting match outcomes in tennis can be categorized in Pre-match prediction variables and point-based predicting variables. Very little research has been done to pre-match predicting variables influencing match outcomes. Boulier and Stekler (1999) were the first to review whether seeding is a good predictive variable of match results in tennis. This turned out to be the case; with the probability of winning a match increasing for the higher ranked player when the difference in rank between the two players increased. The first real

predictive model was made by Klaassen and Magnus (2003). They estimated the win chance of a player at the start of the match through the rank difference between players and updated this prediction based on point outcomes during the game, using match data from Wimbledon between 1992-1995. del Corral (2009) did look for variables favoring upsets[1] in tennis matches. Ranking, and also age of both players turned out to have a significant effect on a match. He also attempted to find the effect of surface and the tournament round on match outcomes, but there were no significant results found adding these variables. del Corral and Prieto-Rodríguez (2010) again checked if ranking difference between players is a good predictor in Grand slam tennis. This again was the case. They also added a lot of other variables, including player specific characteristics like height, preferred hand and age. Only age, previous results in the same tournament, and the lower ranked player being an ex-top 10 player gave significant effects in male Grand slam tennis. Gilsdorf and Sukhatme (2008) added prize money to the equation. With an increase in prize money gap between two tournament rounds, the higher ranked player was more likely to win the match-up.

Kovalchik (2016) reviewed the most relevant articles, including the ones listed before, in the field. She attempted to find the best predictive model so far. In regressive models, a players rank is found to be the main indicator of the probability of winning a match. In the end, she was not able to have a complete model predict outcomes.

What becomes clear from those pre-match predicting models is that most researchers have worked with a small dataset and that they have mainly been using Grand slam tournament matches over a few years. This might be one of the reasons why the impact of several variables is unknown up until now. By using Grand Slam tournaments to do research about tennis matches, there has not been a lot of research about tournament varying data, such as the influence of the amount of tournament rounds. Gilsdorf and Sukhatme (2008) did look into the effect of an increase in rewards in the form of prize money between tournament matches on the outcome of a game, including some lower prestige tournaments. the possible flaw in that research is including Grand slam tournament results.

The ATP does not operate the Grand Slam tournaments (The Australian Open, Roland Garros, Wimbledon and The US Open), as they are organized by the ITF[2], but the ATP does award ranking points based on results in those tournaments and plans other tournaments around them in the tournament calendar. Grand Slams are the only Tournaments on the calendar with a full roster of 7 tournament rounds (128 players). Grand slams should in researches about the differences between tournaments be left out since they are played in a best-of-5 sets format

[1] Upset: the lower ranked player winning
[2] International Tennis Federation

and are played over two weeks. Match results are likely to be influenced in a different way or by other variables than in regular tournaments. They will though influence results slightly, because of the huge impact they potentially have on a players ranking.

In this thesis, research will be done to establish whether draw and tournament type influence win chances of the top seed player. If a single-elimination knockout tennis tournament has more knockout rounds, this could be favorable for higher ranked players. This could, for instance, be due to fitness levels, mental strength, rewards or the experience of having been in that situation before.

This paper only discusses male tennis, as the WTA has a slightly different ranking system[3], and the results in male tennis are expected to be more clear as the top male players tend to dominate the ATP tour more than the top female players do on the WTA tour. Evidence for this is that, in the last decade, only 6 different men divided all the 40 Grand Slam titles between them, while 20 women won one in the past 10 years.

The attempt will be to compare win chances of a tennis player in different size single-elimination tournaments and to add variables influencing win chances to the already existing literature. First, a Probit regression of the variables found in the existing literature will be executed in *Model 1*. Afterwards, the influence of different sized and prestige tournaments on top seed matches will be evaluated in *Model 2*. The influence of byes in certain rounds of a tournament will be analyzed in *Model 3*. The influence of byes, different prestige and different sized tournaments will be tested on the probabilities of the top seed player winning the entire tournament in *Model 4*. Do top seed players have the edge once a tournament has more rounds? Should a tournament director change the format of a tournament to increase playing time of the top seed player?

The Model

This paper attempts to find competitive advantages or disadvantages for the top seed player if the tournament duration or reward varies. The probability of winning a match in a tournament is estimated by doing a Probit regression. A Logit or Probit regression is most fitting, as the probability of winning a match has to be between 0 and 1.

A Probit regression takes on the following shape:

$$\Pr(\mathrm{Y} = 1|\mathrm{X}) = \varphi(X_i * \beta_i + \beta_0)$$

[3] This influences results according to del Corral and Prieto-Rodríguez (2010)

Were Y, being a binary variable, takes on 0 or 1. Pr denotes the probability of Y being 0 or 1. Φ(.) is the cumulative standard normal distribution function. The outcome between brackets is a z-score. The z-score increases with β_i if X_i increases with 1. The probability of the top seed winning a match can be computed from Φ(z).

The Data used is delivered by the ATP. The ATP makes "ATP World Tour tennis data" online available on https://datahub.io/sports-data/atp-world-tour-tennis-data#data under the "creative commons attribution 4.0 international license"[4], which includes: all match data, match stats and tournament stats. All data from 1991 until 2016 will be used unless stated differently, consisting of 3945 ATP singles matches[5] of top seed players, who played a combined 1212 tournaments. Only matches of which both players are in the top 500 at that moment in time[6], tournaments that have a positively listed money prize for the winner[7] and matches with players that have a valid ranking on the day the tournament starts[8] are taken into account. 77.2% of matches is won by the top seed player, which results in a tournament win rate of 25.7% over the same period.

Table 1
Descriptive statistics of results.

	Mean	Std. dev.	Obs
MATCHWIN	.772	.420	3945
TOURNAMENTWIN	.257	.437	1212

Note. The variables take on 0, if a match or tournament is lost. They take on 1, if a tournament or match is won.

1. Model from the existing literature

The variables from the existing literature included in this research are: age, rank rating, prize money and surface. All of these variables, except surface, are found to be significant before. Although not found significant before, surface is assumable to have a significant effect on play.

With age, someone first becomes stronger and fitter becoming an adult, but after a while, nearing the thirties, physical capabilities decrease. As rankings are over the past year, A player could have become physically worse but still be ranked first. This ranking might then give

[4] This license gives the authority to copy and redistribute the data and/or remix and transform the data with no limitation of the final purpose.
[5] The Olympic games, Grand Slams and non-single elimination tournaments are left out, because of non-comparable match/tournament format.
[6] Having to be in the top 500 eliminates 2 observations.
[7] Removes about 500 observations.
[8] Ranking had to be taken over from a different file and this measure is taken for compatibility reasons. A decent amount of masters tournament matches is left out because of this.

wrong information about who actually is the best player in the tournament and decrease chances of the top seed winning. The opposite would be the case being a young player. Del corral and Prieto-Rodriguez (2010) found out that high-ranked male tennis players have a higher chance of losing a tennis match when playing against a younger opponent. A young player, still developing, might not have the ranking reflect the full progress made in the last year. Age is most likely to have a negative impact on win chances. The variable `AGE` will be added to control for the influence of age on match result. Age of the opponent would impact play of the opponent in the same way the top seeds own age does on his own play. Hence, also `AGEOPP` will be added. If the opponents age increases, win chance most likely decreases.

The second variable will be rank rating. The existing literature argues a ranking variable is very important in predicting match outcomes. Del Corral and Prieto-Rodríguez (2010) establish the increase in single match win chance being 1st on the world rankings instead of being 11th is on average around 15% in male grand slam tournaments, but the difference in win chance is about 3% being 51st or being 61st. these results indicate that ranking does not have a linear relationship with win chances. Instead a log base formula of the rank will be used; indicated as `RANKRATING`:

$$10 - log_2\,(rank)$$

A similar formula is proposed by Klaassen and Magnus (2001), assuming tennis quality and seeding are a pyramid[9]. This formula is later used by Del corral (2009) to have a quality indicator in estimating upsets in tennis. The higher the rank (1st is higher than 2nd), the higher the indicator and thus the higher the win chance. The value of the outcome varies from 1 to 10. A positive effect on match results is expected to be found. The same formula will be used on the rank of the opponent to include the effect of the opponents quality on match win chances (`RANKRATINGOPP`). The higher the rank of the opponent and the higher the formula outcome, the lower the win chance of the top seed should be.

Prize money is another variable to be accounted for. Gilsdorf and Sukhatme (2008) found a significance change in the probability of winning a match for the highest ranked player once prize money gap increases. Prize money available in a tournament would be a good variable to indicate the effect of rewards. `PRIZEMONEY` will be added to the model and will indicate the total prize money, in US dollars, available, divided by 100.000. The probability of the top seed winning a match is likely to be positively influenced by the total prize money a tournament has to divide between contenders.

[9] They use $8 - log_2(x)$ as they assess the impact of quality out of 128 ranks, in this research the top500 of the ATP ranking needs to be accounted for.

The final variables being added are dummies for surface. Most tournaments in a year are held on a hardcourt surface (including carpet) followed by clay and a few tournaments are played on grass. Those surfaces differ in characteristics. Du Bois and Heyndels (2007) argue that tennis surface plays a big role on the outcome of a match, as the bounce of the ball and the effect on the ball is different on each surface. Different playstyles are favorable on different courts. For instance: risky play and luck are a more significant factor on faster courts. Sackmann (2015) also addressed that breaking serve is easier on clay (the slowest court). These factors are likely to influence the win chance of the top seed and are thus controlled for. Those variables will be called: DHARD, DCLAY, DGRASS and DCARPET. Win chances of the top seed are expected to be the highest on clay, because luck and risk are less rewarding, but there has not been a significant result on outcome in the earlier named literature.

A summary of the variables is given in Table 2 and a Probit model is executed to estimate the effects of those variables on individual match results in Table 3.

Table 2
Descriptive statistics of variables influencing match results.

	Mean	Std. dev.	Obs
MATCHWIN	.772	.420	3,945
AGE	25.195	3.015	3,945
AGEOPP	25.369	3.559	3,945
RANKRATING	7.834	1.557	3,945
RANKRATINGOPP	4.121	1.297	3,945
PRIZEMONEY	8.672	8.111	3,945
DHARD	.519	.500	3,945
DCLAY	.317	.465	3,945
DCARPET	.104	.205	3,945
DGRASS	.060	.237	3,945

Table 3: effect of variables from the existing literature on match results.

	Model 1	
	Coef.	ME[10]
AGE	-.017**	-.005**
	(.008)	(.002)
AGEOPP	.015**	.004**
	(.007)	(.002)
RANKRATING	.136***	.040***
	(.019)	(.005)
RANKRATINGOPP	-.246***	-.072***
	(.020)	(.006)
PRIZEMONEY	.007**	.002**
	(.004)	(.001)
DCLAY	-.096*	-.028 *
	(.053)	(.016)
DCARPET	-.112	-.033
	(.076)	(.022)
DGRASS	-.188*	-.055*
	(.097)	(.028)
CONSTANT	0.774*** (.297)	
Pseudo-R^2	.048	
Observations	3945	
Log pseudolikelihood	-2017	

*, **, & *** indicate a significance level of respectively 10%, 5% and 1%

The variables from the existing literature, except surface, have a significant influence on match results at a 5% significance level. The age coefficients indicate a similar increase in win chance (around 0.5%) if the opponent becomes a year older as the decrease in win chance if the top seed becomes a year older. The rank rating coefficients show a 4% increase in win chance, *ceteris paribus*, if the rank rating increases with one. This is quite low relative to the impact of the opponents rank rating. If the opponents rank rating increases with one, win chance of the top seed decreases with 7.2%, the other variables remaining constant. Arguably, The top seed, being the highest ranked player in the field, is more favored by having a lower ranked opponent than being higher ranked himself. A tournament prize money increase of $100.000 gives the

[10] ME denotes the average increase in probability if the relevant independent variable increase with 1, *ceteris paribus.*

top seed seemingly more motivation, as this increases the probability of winning a match in the tournament with 0.2%, *ceteris paribus.* Although surface dummies do not have significant results at this level, they still seem to affect match results. Hardcourt is the surface most favorable for the top seed.

The above variables are not the only variables impacting the outcomes of matches. There are lots of in-match variables and psychological variables that have an effect on results, but those are hardly measurable and therefore not implementable. *Model 1* will be a threshold for the upcoming models.

2. How do draw size and tournament type influence the single match result of the top seed player?

A lot of previous papers have only been using Grand slam data, so the influence of tournament types on outcomes has not been discussed yet. Large and small could in this case mean either the draw size differs or rewards differ. Types of tournaments do not have a standard size. Masters tournaments could have 96 players in the draw. The 32 seeded players will then have a bye in the first round and will only have to play 6 matches, while unseeded players have to win 7 matches to win the tournament. A masters tournament could also have 56 or 48 players In the draw. In that case the top 8 seeded or top 16 seeded players get a bye in the first round and only have to win 5 matches to win the tournament. Unseeded or lower seeded players need to win 6. ATP 500 tournament consist of either 32 or 48 players with 8 or 16 seeded players, and ATP 250 tournaments mostly consist of a draw with 28 players of whom 8 are seeded. Masters tournaments have 1000 ranking points available for the winner, ATP 500 events have 500 points available, and an ATP 250 event has 250 points available. An insight in the relationship between draw size and tournament type is given in table 4.

Table 4

Descriptive statistics of draw size means by tournament type.

	Mean	Std. dev.	Obs
ATP250			
DRAWSIZE	32.226	4.542	1,326
ATP500			
DRAWSIZE	37.838	9.285	396
MASTERS			
DRAWSIZE	65.460	15.272	274

Not only tournament types differ in draw size and the amount of ranking points available but also in prize money as shown in table 5.

Table 5

Descriptive statistics of prize money means by tournament type

x100000 in $.

	Mean	Std. dev.	Obs
ATP250			
PRIZEMONEY	5.639	2.217	1,326
ATP500			
PRIZEMONEY	13.390	8.783	396
MASTERS			
PRIZEMONEY	29.067	6.580	274

In this research masters events are the most prestigious tournaments, followed by ATP 500 tournaments and ATP 250 tournaments. masters events have the highest prize pool and reward the most ranking points.

Draw size and tournament type are both used separately although correlation is present. Tournament type is a measure of rewards, and draw size is a measure of the amount of matches to be played. Separate use is necessary, as the data set does not have enough observations of each tournament type in combination with a certain draw size to compare tournament matches of certain tournament type with different draw sizes. In *Model 3,* it is also necessary in evaluating the effect of byes.

ATP restructured their tournaments and the amount of ranking points in 2009. To have a decent amount of observables, “ATP international series” tournaments since 2000 are added to the ATP 250 tournaments, “ATP international series Gold” tournaments since 2000 are added to the ATP 500 tournaments, and “Master series” tournaments are added to ATP Masters Tournaments. All those added matches are the former equivalent of the current events.

Higher rewarding tournaments, in ranking points and prize money, on average, have a larger draw size. Those tournaments have higher rewards for winning the tournament but also in each tournament round. Winning a tournament becomes more important, and each round becomes more valuable. Playing a tournament containing a larger draw also puts more emphasis on Physiology and fatigue. To be able to have some energy left at the end of the

tournament, it would be beneficial to win with a minimum amount of playing time. More effort is put into playing dominantly and not giving the opponent any sort of chance. Average results of matches in tournaments with a certain draw size or certain tournament type are listed below.

Table 6

Descriptive statistics of single match result by draw size and tournament type.

	Mean	Std. dev.	Obs
DRAW28RESULT	.796	.404	225
DRAW32RESULT	.762	.426	2,739
DRAW48RESULT	.786	.411	332
DRAW56RESULT	.783	.413	410
DRAW64RESULT	.811	.393	169
DRAW96RESULT	.829	.380	410
ATP250RESULT	.766	.423	1,326
ATP500RESULT	.801	.400	396
MASTERSRESULT	.818	.387	274

As shown in Table 6, Match results per draw size indicate that the relative amount of single matches won increases if a tournament has a bigger draw. Mean match results per tournament type indicate that tournament types are likely to have a positive effect on single match outcomes as well, with the probability of the top seed player winning a match increasing if the tournament has more prestige.

Average prize money increases as well if the tournament type changes from ATP 250 to ATP 500 and from ATP 500 to a masters event. The increase in win chance per tournament type could be due to the increase in prize money. The increase in prize money is larger than the increase in win chance. The net[11] effect of draw size and tournament type on single match win chance is therefore likely to be found negative.

A similar regression will now be used as in M*odel 1*, but different draw sizes and different types of tournaments will be added. First, only the effect of an increase in draw size and change of tournament type will be tested. Afterwards, there will be controlled for the other influenceable

[11] Effect when accounting for prize money differences.

variables. Note that the amount of observations will be lower when comparing the types of tournaments, because only tournaments from the year 2000 onwards are taken into account.

Table 7

Probit results of the influence of draw size and tournament type on match results.

	Model 2							
	Coef.	ME	Coef.	ME	Coef.	ME	Coef.	ME
DDRAW32	-.112	-.034	-.223**	-.045**				
	(.098)	(.030)	(.106)	(.031)				
DDRAW48	-.033	-.010	-.245*	-.075*				
	(.122)	(.037)	(.135)	(.040)				
DDRAW56	-.044	-.013	-.340**	-.100**				
	(.117)	(.035)	(.134)	(.039)				
DDRAW64	.054	.016	-.351**	-.103**				
	(.146)	(.044)	(.176)	(.051)				
DDRAW96	.123	.037	-.394*	-.116*				
	(.201)	(.061)	(.236)	(.069)				
ATP500					.117	.035	-.141	-.40
					(.081)	(.024)	(.102)	(.029)
MASTERS					.180*	.053*	-.501***	-.143***
					(.096)	(.028)	(.191)	(.055)
AGE			-.022***	-.006***			-.014	-.004
			(.008)	(.002)			(.010)	(.003)
AGEOPP			.013*	.004*			.016*	.004*
			(.007)	(.002)			(.009)	(.003)
RANKRATING			.147***	.043***			.171***	.049***
			(.020)	(.006)			(.030)	(.008)
RANKRATINGOPP			-.248***	-.073***			-.224***	-.064***
			(.020)	(.006)			(.028)	(.008)
PRIZEMONEY			.011***	.003***			.023***	.007***
			(.004)	(.001)			(.007)	(.002)
DCLAY			-.075	-.022			-.083	-.024
			(.055)	(.016)			(.075)	(.021)
DCARPET			-.108	-.032			-.271*	-.077*
			(.078)	(.023)			(.158)	(.045)
DGRASS			-.156	-.046			-.192	-.055
			(.098)	(.029)			(.139)	(.040)

CONSTANT	.826*** (.095)	1.054*** (.332)	.726*** (.038)	.264 (.444)
Pseudo-R^2	.001	.050	.002	.052
Observations	3945	3945	1996	1996
Log pseudolikelihood	-2117	-2014	-1049	-997

*, **, & *** indicate a significance level of respectively 10%, 5% and 1%

At first sight, there is nearly no significant difference between win chances in different size tournaments. Only masters tournaments provide an 5% increase in win chance, at a 10% significance level, in comparison with ATP 250 tournaments *ceteris paribus*.

The outcome chances drastically controlling for the earlier found match result influencing variables. Instead of a 5% increase in win chance if the tournament type changes from ATP 250 to masters, a 14% decrease is found. The draw size dummies all become significant at a 10% significance level, with the coefficient becoming more negative if draw size increases. The larger the draw, the smaller the win chance of a single match seems to become. The outcomes with control variables are more inclusive, as the Pseudo-R^2 increases a lot.

The change in coefficients of tournament type dummies and draw size dummies has to be because of prize money. All existing variables except prize money have similar coefficients in *Model 1* as in *Model 2*. The prize money coefficient, on the other hand, more than doubles in this model with dummy variables for draw size and tournament type in comparison with *Model 1*. To evaluate the impact of prize money, another Probit model will be made without it.

Table 8

A copy of *Model 2* excluding prize money.

	Model 2 continued			
	Coef.	ME	Coef.	ME
DDRAW32	-.220**	-.056**		
	(.106)	(.031)		
DDRAW48	-.193	-.057		
	(.133)	(.039)		
DDRAW56	-.271**	-.080**		
	(.131)	(.038)		
DDRAW64	-.175	-.051		
	(.165)	(.049)		
DDRAW96	-.177	-.052		
	(.223)	(.066)		

ATP500			-.001	-.000
			(.091)	(.026)
MASTERS			-.016	-.005
			(.126)	(.036)
AGE	-.019**	-.006**	-.009	-.003
	(.008)	(.002)	(.010)	(.003)
AGEOPP	.014**	.004**	.018**	.005**
	(.007)	(.002)	(.009)	(.003)
RANKRATING	.159***	.047***	.186***	.053***
	(.019)	(.006)	(.029)	(.008)
RANKRATINGOPP	-.239***	-.070***	-.213***	-.061***
	(.019)	(.006)	(.028)	(.008)
DCLAY	-.091*	-.027*	-.123*	-.035*
	(.055)	(.016)	(.074)	(.021)
DCARPET	-.117	-.033	-.230	-.066
	(.078)	(.023)	(.154)	(.045)
DGRASS	-.177	-.052	-.182	-.052
	(.098)	(.029)	(.139)	(.040)
CONSTANT	.893*** (.321)		.060 (.434)	
Pseudo-R^2	.049		.048	
Observations	3945		1996	
Log pseudolikelihood	-2017		-1001	

*, **, & *** indicate a significance level of respectively 10%, 5% and 1%

Removing prize money makes the coefficients of tournament type and most coefficients of draw size very insignificant. This indicates that prize money is a crucial factor in determining the impact of draw size and tournament type on the pre-match win chance of the top seed player. If other variables are unchanged, the top seed player is less likely to win a single match if the draw size increases or a tournament becomes more prestigious, but because of the consequential increase in prize money, the probability of winning in a single match in a more prestigious tournament rises.

3. How does a bye affect win chances of the top seed player?

Nowadays, more than halve of the ATP tour tournaments have byes involved. Including byes in a draw is an instrument used by tournament directors. Byes provide the top seed players a guaranteed spot in the second round. This makes tournaments more attractive for higher rated players, because a guaranteed second round result earns a higher guaranteed amount of prize

money, more guaranteed ranking points and, even more important because of the demanding tennis calendar, a bit more rest between tournaments.

A bye is rewarded to a high seeded player if a roster of 32, 64 or 128 players is not fully filled. A draw of 28 players would leave four players without an opponent in the first round. The four highest ranked players will immediately proceed into the second round, called a bye. Because of draw size determining whether the top seed has a bye[12], the influence of byes in a model with draw size cannot be evaluated.

Jeff Sackmann (2012) analyzed which players profit from byes in a tournament. Players are placed into three groups: seeded players with a bye, seeded players without a bye and unseeded players. Seeded players with bye are mainly favored by not having to play the first round. This way they do not face the risk of getting eliminated in that round. A countervailing effect would be the other seeded players with a draw having the same advantage and therefore are more likely to be future opponents. As seeded players are the highest ranked, most skillful players, this slightly decreases the chance of winning the tournament. This negative effect of being granted a bye is smaller than the benefits of having a bye for those players. Seeded players without a bye, on the contrary, are losers of byes implemented in a draw. They do not have the benefits of skipping the first round and are more likely to face seeded players with a bye in the final stages of the tournament. Unseeded players are slightly benefitted by byes. They are not expected to last a lot of rounds or be a real contender for the title. The main concern for them at the start of a tournament is to win the first round. With byes added to the draw, they cannot face the top few players in the first round and consequently their win chance in the first round increases.

Following The findings of Sackmann (2012), byes do provide the top seed player with a 100% win rate in the first round, and byes decrease win chances in the last few rounds[13]. They would not really affect the rounds in between. This does not have to be the case, as the analysis does not account for any physical or mental state. Having another day or two of rest could benefit the top seed player as well. This effect will most likely be the greatest in the 2nd round but might also percolate into the third and fourth round.

Because effects of byes are assumed to be the greatest in round 2 and in the final, those rounds will be the dependent variables. Those rounds are also, not unimportantly, two rounds certain to be played in a tournament. As the final could be round 5 of a tournament, adding round 5 would be interfering with final results.

[12] The top seed never has a bye when the draw size is 32, 64 or 128 and has a bye in all other cases.
[13] The amount of tournament rounds with decreasing win chance is dependent on the amount of other seeded players with a bye at the start of the tournament.

Table 9 shows quite a small different in means between round 2 results with and without bye, but the amount of matches won with is slightly (1.9%) higher than without.

Table 9

Descriptive statistics of round 2 results, with and without bye.

	Mean	Std. dev.	Obs
ROUND2RESULT			
WITH BYE	.839	.368	310
ROUND2RESULT			
WITHOUT BYE	.820	.385	694

The difference in means is quite small, and the amount of observation might be too little to find a significant effect of playing the 2nd round with or without bye. If a significant result is found, it is likely that the additional days of rest influence the probability of winning the second round match in a positive way.

Following Sackmann (2012), A bye should have negative impact on a top seed player in a final, because the chance of facing a higher rated player becomes larger. The summarizing statistics in table 10 show the contrary. If the top seed had a bye, the amount of matches won was 10% higher.

A Probit regression of the 2nd round results and finals results is executed and the results are shown in table 11.

Table 10

Descriptive statistics of results in a final, with and without bye.

	Mean	Std. dev.	Obs
FINALRESULT			
WITH BYE	.776	.419	125
FINALRESULT			
WITHOUT BYE	.673	.470	318

Table 11

Probit results of the influence of byes on 2nd round and final round match results.

	Model 3							
	Probit of round 2 results				Probit of finals results			
	Coef.	ME	Coef.	ME	Coef.	ME	Coef.	ME
DBYE	.074	.019	-.146	-.036	.311**	.107**	368**	.124**
	(.102)	(.026)	(115)	(.028)	(.145)	(.050)	(.157)	(0.53)
AGE			-.020	-.005			-.035	-.012
			(.016)	(.004)			(.023)	(.008)
AGEOPP			.007	.002			.015	.005
			(.014)	(.003)			(.020)	(.007)
RANKRATING			.124***	.030***			.209***	.070***
			(.039)	(.010)			(.062)	(.021)
RANKRATINGOPP			-.322***	-.079***			-.300***	-.101***
			(.068)	(.016)			(.055)	(.018)
PRIZEMONEY			.009	.002			.019*	.006*
			(.008)	(.002)			(.011)	(.004)
DCLAY			-.235**	-.058**			.119	.040
			(.111)	(.027)			(.159)	(.053)
DCARPET			-.276*	-.068*			.050	.017
			(.157)	(.038)			(.235)	(.079)
DGRASS			-.398**	-.097**			.046	.015
			(.193)	(.047)			(.300)	(.101)
CONSTANT	.915*** (.056)		1.581*** (.624)		.448*** (.073)		.705 (.889)	
Pseudo-R^2	.001		.045		.009		.096	
Observations	1004		1004		443		443	
Log pseudolikelihood	-464		-444		-267		-244	

*, **, & *** indicate a significance level of respectively 10%, 5% and 1%

A bye has not been found to have significant impact on second rounds results. Quite noteworthy is the sudden significant impact of the surface dummies in the second round of a tournament. Hardcourt again being most favorable for the top seed.

Even though the amount of observations is limited, byes have a positive impact on the outcomes of finals at a 5% significance level. Contrasting with previous literature, this effect is positive. When reaching the final, the win chance of the top seed is 10% higher when the 1st round was a bye, *ceteris paribus*. This is possibly because of earlier named rest periods

and having more energy left. The surprising part about this theory is the missing significance in round 2, where the relative rest surplus of the top seed should be the highest, because the opponent has already played a match that week. The physical part of tennis needs further research to be able to be implemented.

Rank rating, as well, has an increased influence on final results in comparison with the influence on round 2 results and in comparison with outcomes found in *Model 1*. This in what would be expected to be found following Gilsdorf and Sukhatme (2008). If the price gap between rounds increases, higher ranked players have an increased win probability. The highest absolute price gap between rounds is from semi-final to final, so the higher ranked the player the larger the increase in win chance from round to round.

4. How do draw, byes and tournament type influence the win chance of the entire tournament?

In the end, the tournament director wants the top seed to win the tournament. Having a renowned name win the tournament, makes the tournament more memorable and raises publicity. Increasing tournament length or adding byes would be easy implementable features.

Similar variables are used as in the previous few questions. The dependent variable is the tournament win chance. The tournament win chance takes on a 1 if a final in the dataset is won and takes on a 0 for each loss in the dataset. In *Model 4*, the variables accounting for the quality and age of the opponent are left out, because multiple opponents are faced in a tournament. The data set will again be smaller comparing between tournament types, because only tournaments from 2000 onwards are observed. A bye will be added in the Probit model with tournament types, as having a bye is a consequence of draw size, but it is not a consequence of tournament type.

Simply looking at the amount of matches on the road to becoming champion, win chance should be the greatest in tournaments with a draw size of 28, because only 4 matches have to be won to win the tournament. In draws with 32 to 63 players, 5 matches have to be won by the top seed. In draws with 64 to 127 players, 6 matches have to be won. Not only the amount of matches plays a role. Prize money proved an important factor and is correlated with prize money, which could have a higher win chance in larger draws as a consequence. Tournament type is more correlated with rewards than with draw size, but part of the rewards is controlled for by adding prize money. Win chance is expected to be higher in masters than in ATP 500 events and higher in ATP 500 events than in ATP 250 events.

Table 12 gives an overview of mean tournament results of the top seed player.

Table 12

Descriptive statistics of tournament result by draw size and tournament type.

	Mean	Std. dev.	Obs
DRAW28RESULT	.387	.490	75
DRAW32RESULT	.235	.424	851
DRAW48RESULT	.283	.453	99
DRAW56RESULT	.276	.449	123
DRAW64RESULT	.304	.465	46
DRAW96RESULT	.333	.485	18
ATP250RESULT	.265	.442	422
ATP500RESULT	.295	.458	112
MASTERSRESULT	.324	.471	74

Most variables are likely to have the same effect on tournaments as on matches, but byes are more likely to improve win chances of a tournament than of a single match. Again byes cannot be added to the model including draw sizes, as byes are a direct consequence of the draw size. A Probit model of tournament results gives the following output:

Table 13

Influence of draw and tournament type on the tournament results of the top seed player.

	Model 4			
	Coef.	ME	Coef.	ME
DDRAW32	-.635***	-.201***		
	(.166)	(.053)		
DDRAW48	-.699***	-.222***		
	(.223)	(.071)		
DDRAW56	-.810***	-.257***		
	(.220)	(.070)		
DDRAW64	-.868***	-.275***		
	(.308)	(.098)		
DDRAW96	-.804**	-.255**		
	(.393)	(.124)		

ATP500			-.420**	-.138**
			(.179)	(.059)
MASTERS			-1.134***	-.372***
			(.353)	(.115)
DBYE			.340***	.112***
			(130)	(.042)
AGE	-.037***	-.012***	-.035**	-.011**
	(.014)	(.004)	(.018)	(.006)
RANKRATING	.153***	.048***	.197***	.065***
	(.033)	(.011)	(.047)	(.016)
PRIZEMONEY	.002	.001	.028**	.009**
	(.008)	(.003)	(.012)	(.004)
DCLAY	.027	.008	-.013	-.004
	(.095)	(.030)	(.132)	(.043)
DCARPET	-.088	-.028	-.380	-.125
	(.139)	(.044)	(.290)	(.095)
DGRASS	-.061	-.019	-.149	-.049
	(.173)	(.055)	(.247)	(.081)
CONSTANT	-.295 (.483)		-1.369** (.644)	
Pseudo-R^2	.032		.056	
Observations	1212		608	
Log pseudolikelihood	-668		-339	

*, **, & *** indicate a significance level of respectively 10%, 5% and 1%

In evaluating draw size in *Model 4*, the most striking finding is that prize money nearly becomes unimportant. It seems that the amount of matches played, indeed, is the most important factor in tournament win chance. The larger draw size becomes, the lower the tournament win chance. Age and rank are found significant again.

When assessing tournament type, similar results are found as in *Model 2*. Whilst the top seed wins more masters tournaments than smaller tournaments, the coefficients are negative and indicate a lower win chance. The win chance of an ATP 250 tournament is 37.2% higher than a masters tournament if other variables are constant. As found before, prize money plays a role in this phenomenon. Masters tournaments have considerably more prize money available, and the tournament win chance of a top seed player increases on average with 0.9% for each additional $100.000. Byes are more present in more prestigious

tournaments, as found in table 14, and play a significant role, with an 11.2% raise in win chance, *ceteris paribus*, If the top seed has a bye.

Table 14

Descriptive statistics of byes by tournament type.

	Mean	Std. dev.	Obs
ATP250	.216	.412	1,326
ATP500	.295	.457	396
MASTERS	.522	.500	274

Summary

This research attempted to find the influence of tournament format, or more specifically the influence of draw size, tournament type and byes on individual match result and on tournament outcomes. Four Probit models were used to establish signs and significance of those variables, while also variables found significant in previous literature were added.

The main findings concerning draw are a significant decrease in win chance in matches and tournaments if draw size increases from 28 to a larger draw size. It is not clear whether coefficients of other draw sizes significantly differ from each other, but this could be expected due to the coefficients becoming more negative with the draw size becoming larger. The tournament win probability in a tournament with a larger draw is found to be smaller. Likely, because the amount of matches that has to be won to win the tournament is higher. The tournament win chance decreases with more than 20% once the top seed player has to play more than 4 matches.

In tournament type win results, there is a significant decrease in single match result when playing in a masters event instead of in an ATP 250 tournament. Tournament results shows a negative change when being top seed in an ATP 500 event or masters event in comparison with an ATP 250 event. Main factors behind this result are byes (in the tournament type models) and prize money. More prestigious tournaments have an increase in the amount of byes. As seen in *Model 4*, Byes increase the tournament win chance significantly. Where prize money has a moderate effect on match outcomes in tournaments with different draw sizes, it

is a larger factor in different types of tournaments. It has to be seen as the main variable of interest for top seed players to put effort into a match.

Discussion and implementation.

In concluding results of this paper, a few things have to be considered. In this research results from 1991 until 2016 have been evaluated to get to a decent amount of observations, but time varying variables are not taken into account. It is likely that prize money has increased over time and byes have mainly been applied recently. This is a downfall from not using Grand slam data and should be researched further. Secondly, all top seeds are assumed to have the same level of importance in this research. In real life this is not the case. Having one player who is considerably better than others, makes that player far more important than the top seed if the top few seeds are ranked 15, 16 and 17 in the world.

As mentioned before, The top-seed player is more often than not the eyecatcher of a tournament. He is the player people come to watch and see. He should play as many matches as possible and preferably reach the finals to generate the biggest revenue for organizers. A large factor for a top seed player to perform is prize money, but this would higher the costs and lower profits. A tournament director should evaluate the costs and revenues to make a decision. Draw size should neither be increased. This only leads to lower match win chance and a lower probability of winning the tournament. A good instrument for a tournament director to use would be including byes in the tournament. A tournament director could apply a draw of 56 instead of 64, or 28 instead of 32. The likelihood of the top seed winning the tournament increases this way. the number of first round matches is decreased, which reduces income, but those matches have the lowest interest, as the top seed would win quite easily most of the times. Hence, this reduction of income is most likely outweighed by the expected pay-offs.

This research cannot only be applied in the sports industry, but it could also be applied in intrafirm promotion tournaments. This research found comparable results as Lazear and Rosen (1981) with prize money being a driving factor in a tournament. In addition, implementing byes could be a good instrument for directors to have preferred participants win an infirm promotion tournament. Giving an exciting talented prospect a chance in a higher ranked and payed job in a firm, without having to compete with other colleagues for that job, could help in the development of that prospect and get that person as high as possible in the hierarchy. This has to be further researched to be conclusive.

The results could, possibly, taking the differences between sports, professionality, and genders, be applied to different sports or female/amateur tennis. Even at amateur level, having the number one seed playing multiple matches could increase bar revenues and publicity[14].

[14] Publicity could improve reputation and increase registrations or attendance in upcoming years.

Appendix

Introduction to the ATP

The Association of Tennis Professionals (ATP) is the organizer of all events on the ATP tour (former ATP World tour). Those events are exclusively for male players, as the WTA organizes the big tournaments for women.

The ATP organizes different tournaments on different levels. The ATP tour consists of several single-elimination knockout tournaments: ATP challenger tour (80-125), ATP 250, ATP 500 and ATP masters 1000 tournaments. The numbers represent the amount of ranking points awarded to the winner. Winning a Grand Slam earns 2000 ranking points. The total amount of points received in tournaments in the past year accumulates to a spot on the global rankings. The results of a tournament stand until the same week in the upcoming year, so a ranking consists of the ranking points achieved in the past 52 weeks.

The ATP also (partly) organizes the ATP-finals, Davis cup and the Laver cup. Those tournaments are not taken into account in this research, as they are not played in a single-elimination knockout tournament format.

References

Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors?: an evaluation. *International Journal of Forecasting*, *15*(1), 83–91. https://doi.org/10.1016/s0169-2070(98)00067-3

Chmait, N., Robertson, S., Westerbeek, H., Eime, R., Sellitto, C., & Reid, M. (2020). Tennis superstars: the relationship between star status and demand for tickets. *Sport Management Review*, *23*(2), 330–347. https://doi-org.eur.idm.oclc.org/10.1016/j.smr.2019.03.006

del Corral, J. (2009). Competitive balance and match uncertainty in grand-slam tennis. *Journal of Sports Economics*, *10*(6), 563–581. https://doi.org/10.1177/1527002509334650

del Corral, J., & Prieto-Rodríguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting*, *26*(3), 551–563. https://doi.org/10.1016/j.ijforecast.2009.12.006

DeVaro, J. (2006). Internal promotion competitions in firms. *The RAND Journal of Economics*, 37, 521-542. doi:10.1111/j.1756-2171.2006.tb00029.x

Du Bois, C., & Heyndels, B. (2007). It's a different game you go to watch: competitive balance in men's and women's tennis. *European Sport Management Quarterly*, *7*(2), 167–185. https://doi-org.eur.idm.oclc.org/10.1080/16184740701353349

Gilsdorf, K. F., & Sukhatme, V. A. (2008). Testing Rosen's Sequential Elimination Tournament Model: Incentives and Player Performance in Professional Tennis. *Journal of Sports Economics*, *9*(3), 287–303. https://doi.org/10.1177/1527002507306790

Klaassen, F. J. G. M., & Magnus, J. R. (2001). Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model. *Journal of the American Statistical Association*, *96*(454), 500–509. https://doi.org/10.1198/016214501753168217

Klaassen, F. J. G. M., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, *148*(2), 257–267. https://doi.org/10.1016/s0377-2217(02)00682-3

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, *12*(3), 127–138. https://doi.org/10.1515/jqas-2015-0059

Lazear, E. P., & Rosen, S. (1979). *Rank-order tournaments as optimum labor contracts* (Ser. Working paper - national bureau of economic research, no. 401). National Bureau of Economic Research. https://www-nber-org.eur.idm.oclc.org/papers/w0401.pdf

Rijnmond. (2020, January 9). *Krajicek: 'Deelnemersveld in Ahoy precies zoals we willen'*. Retrieved from: https://www.rijnmond.nl/nieuws/190636/Krajicek-Deelnemersveld-in-Ahoy-precies-zoals-we-willen

Rosen, S. (1986). Prizes and Incentives in Elimination Tournaments. *The American Economic Review, 76*(4), 701-715. Retrieved May 10, 2020, from www.jstor.org/stable/1806068

Sackmann, J. (2012). *Who Benefits From Byes?* Heavy Topspin. http://www.tennisabstract.com/blog/2012/07/27/who-benefits-from-byes/

Sackmann, J. (2015). *The Case for Novak Djokovic … and Roger Federer … and Rafael Nadal* Heavy Topspin. http://www.tennisabstract.com/blog/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/