

Social Composition & Sport Dropout

A study on Socioeconomic and Demographic Reasons for Soccer Club Dropout

Leydi Johana Breuls



Universiteit Utrecht

Social Composition & Sport Dropout

A Study on Socioeconomic and Demographic Reasons for Soccer Club Dropout

Utrecht University

Utrecht School of Governance

Research Master Public Administration and Organizational Science

Supervisors:

Prof. dr. Maarten van Bottenburg

Prof. dr. mr. Mark Bovens

Student: Leydi Johana Breuls

Student number: 3340481

Contact: Leydi.Breuls@gmail.com



Universiteit Utrecht

Preface and acknowledgements

The initial idea to do a research concerning sports was based on a PhD opportunity given to research master students of the Utrecht School of Governance. During the following one-and-a-half years I was engaged in finding a research question that was scientifically interesting as well as interesting for praxis, gathering the right data to answer this question, analyzing the data using practical and scientifically sound means, and reporting this. The product of this process is in the pages that follow.

This research is part of a larger research project initially called “the end of membership?” This larger research project tries to achieve a better understanding of the changes in membership and membership forms of different sports in the Netherlands. Inspired by this overarching theme, I decided to research sport club dropout. The availability of data of the KNVB helped to conduct research on dropout in soccer clubs in the Netherlands.

I sincerely hope that the reader will recognize the value of this research for a scientific understanding of what sport dropout is and how this is related to the social composition of the sport club, as well as the value this research has for praxis. For those people working with sport clubs on a day-to-day basis, whether as a coworker in a sport association, or as board member of a sport club, I have written an executive summary and a list of recommendations, which can be found on page 5.

This research would not have been possible without the support of a number people. First of all, I would like to thank the Royal Dutch Football Association (KNVB) for the access to their membership data; without this data this research would not have been possible. Special thanks to Laura Jonkers and Jan Kok of the KNVB, who helped me retrieve the right data. Secondly, I am indebted to Maarten van Bottenburg and Mark Bovens, my supervisors: Thanks to their advice and support this research has grown, and so have I!

There are a number of people I would like to thank as well: Karin and Aad for giving me valuable statistics advice, and being patient with me. Thanks Marc for hearing me out day

4 | Social Composition & Sport Dropout

after day, while struggling with getting everything analyzed and on paper. Thanks to my parents, who always support me – both in encouragements and in-kind. Thanks to Vincent, to whom I am indebted for his unconditional support, but also for his advice on research intelligence.

Leydi Johana Breuls

Amersfoort

October 31, 2014

Executive summary

Issue

Dropout numbers are estimated at 130.000 people per year by NOC*NSF, which should cover all sports registered at NOC*NSF. However, this study shows that dropout at the KNVB alone exceeds this estimation. This discloses that dropout is an underestimated issue that requires more attention from sport associations and sport clubs. Sport associations and sport clubs are more inclined to focus on membership recruitment than on retaining members. Still, membership recruitment is much more expensive than retaining members. The question then is, why do people dropout of a club? (and what can we do about it?) Individual motivations for dropout are much more researched to date than club aspects that could motivate individuals to dropout. Consequently, this research looks into sport club aspects for dropout, focusing on the social composition of the sport clubs and their dropout rates. The research question is: *To what extent does the social composition of a sport club play a role in the decision of members to end the membership?*

Purpose

The purpose of this research is to understand better what the influence of social composition is on people's decision to end their membership. To answer the research question, this research looked into ethnicity, income, education level, gender, and age of all members of all soccer clubs, and related this to the dropout rate of that sport club. This understanding of sport club reasons for dropout leads a number of recommendations for soccer clubs, the KNVB, and in extension, other sports. These recommendations can be taken into account when formulating ways of tackling dropout in general, or in specific groups (of clubs), as dropout is a major factor in membership loss.

Approach

This research draws upon the theoretical perspective of Putnam, especially the idea of hunkering down. Hunkering down points out that people (literally) tend to look down when social contexts become too complex. This could also indicate why people dropout of sports. We expected from theory that non-Westerners (ethnicity), low minimum income, and low

6 | Social Composition & Sport Dropout

education levels led to fewer dropouts, and that for gender and age specific conditions were required. The question was answered using KNVB and Statistics Netherlands data. The analyses done were correlations and multiple regression.

Results

From the table below it can be derived that non-Western ethnicity, males, high maximum age, and high minimum age have highest correlation percentages with dropout. This indicates that when either one of these variables or dropout goes up, the other variable goes up as well. The problem with establishing correlations is that it does not represent causality. However, these results can be used as an initial idea of how social composition and dropout are correlated.

Independent (high values of)	↔	Outcome
Western ethnicity	-17.64%	Dropout
Non-Western ethnicity	17.64%	
Minimum income	-.85%	
Low education	.76%	
Medium education	4%	
High education	3.3%	
Male	8.5%	
Female	-8.5%	
Minimum age	5.95%	
Average age	2.79%	
High age	6.71%	

The results of multiple regression were a model that included all the aspects of social composition that were deemed influential on dropout: non-Western ethnicity, income, being highly educated, and age. The model explained about 30% of the dropout in soccer clubs. These aspects of social composition need thus be taken into account when trying to tackle dropout in soccer clubs, and perhaps in sport clubs in general (see table below).

Independent (high values of)	→	Outcome
Non-Western ethnicity	.470	Dropout
Minimum income	-.052	
Average income	-.109	
Maximum income	.135	
High education	.055	
Average age	-.063	
Maximum age	-.066	

Recommendations

There are a few recommendations that can be formulated looking at the research outcomes of the current study.

1. First of all, it is important as a sport association and as a sport club to keep track of dropout. This makes developing precise measures to tackle dropout easier. Dropout is worth the attention as it is a major source of member loss, even when taking into account new memberships (see conclusion.)
2. Dropout does not adhere to the theory of sport participation. One of the key implications is that tackling dropout should be related to current members and past records of who drops out. The current study indicated that more non-Westerners in a club leads to more dropout and that higher levels of highly educated people leads to more dropout. These two aspects should be researched more in depth in order to create association and/or sport club wide policies on how to tackle dropout due to these groups.
3. The KNVB (and other sport associations) would profit from using research and/or business intelligence programs to make trends (like dropout) more easily interpretable in order to create policies that tackle the major issues the sport association faces. This also creates an opportunity to respond to predicted events before they actually happen.

Table of Contents

Preface and acknowledgements	3
Executive summary	5
Issue.....	5
Purpose.....	5
Approach.....	5
Results.....	6
Recommendations.....	7
1. Introduction	11
Outline.....	14
2. Conceptual clarification	15
2.1 Sport & society.....	15
2.2 Hunkering down.....	16
2.3 Features of the sport club.....	17
2.4 Sport club membership.....	18
2.5 Social composition.....	19
2.6 Dropout.....	20
3. Theoretical framework	21
3.1 Sport dropout & ethnicity.....	22
3.2 Sport dropout & income.....	23
3.3 Sport dropout & education level.....	23
3.4 Sport dropout & gender.....	24
3.5 Sport dropout & age.....	25
3.6 Theoretical model.....	27
4. Data and Methods	28
4.1 Data description.....	28
4.1.1 Dataset Statistics Netherlands – Postal codes and ethnicity.....	28
4.1.2 Dataset Statistics Netherlands – Postal codes and income.....	28
4.1.3 Dataset Statistics Netherlands – Postal codes and education level.....	29
4.1.4 Dataset Royal Dutch Football Association (KNVB) – Member information.....	29
4.1.5 Dataset Royal Dutch Football Association (KNVB) – Club information.....	30
4.1.6 Dataset Royal Dutch Football Association (KNVB) – Club postal codes.....	30
4.2 Methodology.....	31
4.2.1 Research Intelligence.....	31
4.2.2 Statistical analyses.....	32
4.2.3 Methodological reflection.....	34
4.3 Operationalization & measurements.....	35
4.3.1 Ethnicity.....	35
4.3.2 Income.....	35
4.3.3 Education level.....	35
4.3.4 Gender.....	36
4.3.5 Age.....	36
4.3.6 Dropout.....	36
5. Descriptive statistics	37
5.1 Ethnicity.....	37
5.2 Income.....	38
5.3 Education level.....	39

5.4 Gender	40
5.5 Age	41
5.6 Members & dropout	41
5.6.1 Dropout	43
6. Analyses & Results	45
6.1 Dropout & independent variables I	46
6.1.1 Dropout & ethnicity	46
6.1.2 Dropout & income	47
6.1.3 Dropout & education	48
6.1.4 Dropout & gender	49
6.1.5 Dropout & age	50
6.1.6 Subconclusion	51
6.2 Dropout & independent variables II	51
6.2.1 Dropout & ethnicity – controlled for education	52
6.2.3 Dropout & income – controlled for education	53
6.2.4 Dropout & gender – controlled for age	53
6.2.5 Subconclusion	54
6.3 Conclusion of correlations	55
6.4 Dropout & social composition	56
6.4.1 Summary of the model	56
6.4.2 Model parameters	57
6.4.3 Independent variables' contribution to the model	59
6.4.4 Standard deviation change in independent variables and dropout	60
6.4.5 Extreme cases	62
6.5 Conclusions of social composition model	63
7. Conclusion	65
7.1 Dropout & ethnicity	67
7.2 Dropout & income	68
7.3 Dropout & education level	68
7.4 Dropout & gender	69
7.5 Dropout & age	69
7.6 Correlations & multiple regression	70
7.7 Dropout & social composition	71
8. Discussion	72
8.1 Conceptual discussion	72
8.2 Methodological discussion	73
8.3 Discussion of results	75
8.4 Recommendations for future research	76
8.5 Possible underlying mechanisms	77
8.6 Homogeneity and heterogeneity	78
9. Encore	80
9.1 Dropout & indexes I	80
9.2 Dropout & indexes II	82
9.3 Dropout & social composition indexes	84
9.3.1 Summary of the model	84
9.3.2 Model parameters	85
9.3.3 Indexes' contribution of the model	87
9.3.4 Standard deviation change in indexes and dropout	87
9.3.5 Extreme cases	88
9.3.6 Checking assumptions	88
9.4 Conclusions of ethnicity and gender indexes-based model	90

Literature	91
Appendices	95
1. QlikView model	95
2. Correlations - output.....	96
<i>a. Bivariate.....</i>	<i>96</i>
<i>b. Partial - bootstrapped</i>	<i>101</i>
3. Multiple regression – full description.....	111
<i>a. Descriptives.....</i>	<i>111</i>
<i>b. Summary of the model.....</i>	<i>111</i>
<i>c. Model parameters.....</i>	<i>113</i>
<i>d. Excluded variables.....</i>	<i>118</i>
<i>e. Assessing the assumption of multicollinearity.....</i>	<i>119</i>
<i>f. Casewise diagnostics</i>	<i>120</i>
<i>g. Checking assumptions.....</i>	<i>122</i>
4. Multiple regression - output.....	127
<i>a. Correlations table.....</i>	<i>127</i>
<i>b. Model summary</i>	<i>130</i>
<i>c. ANOVA</i>	<i>131</i>
<i>d. Model parameters.....</i>	<i>132</i>
<i>e. Excluded variables.....</i>	<i>136</i>
<i>f. Casewise diagnostics</i>	<i>137</i>
5. Factor analysis.....	142
<i>Simple regression with education.....</i>	<i>143</i>
6. Multiple regression new model.....	145
8. Encore - output.....	147
9. Reflection and philosophy of science.....	155

1. Introduction

August 13, 2014 – Sport and strategy – “When it comes to attracting new members, team sports like soccer and hockey are still going strong,” says Jan-Willem van de Roest, PhD researcher at Utrecht University. “However, other sports, like baseball, should focus on retaining members” (Barreveld, 2014).

On a yearly basis about 130.000 people quit playing sports at a sport club according to the *Sportersmonitor 2012* (Hendriksen & Hoogwerf, 2013).

In recent years, the idea of Putnam (2001) that there is a decline of community has reached Dutch discourse on (social) participation. In Putnam’s work voluntary associations are important, and he gives them a central role in civil society. However, there are no direct indications that such a decline of community thesis exists in the Netherlands (Pharr, Putnam, & Dalton, 2000). Still, other work of Putnam (2007) addresses the idea of hunkering down, which points at individuals literally looking down due to increasing social complexity.

This research studies an interesting type of voluntary association: the sport club. The role of sports in society is discussed by numerous scholars (see e.g. Bourdieu, 1986; Putnam, 2001; Siisiäinen, 2000; Tsai & Gracy, 1976; van Sterkenburg, 2012). Sport is believed to contribute to social capital, social cohesion, and better health. This research taps into social composition aspects of sport clubs and tries to connect this with dropout at the sport club level, in order to create a better understanding of why people dropout of sports.

Dropout is a topic in sports that is mostly researched in the adolescent age group or at the top sport level, and taps into personal and individual reasons for sport membership termination, much less is known about other factors that could contribute to dropout (Collard & Hoekman, 2013). Motivations to quit playing organized sports are roughly defined by age category, with time, money, and physical constraints being primary reasons (Hendriksen & Hoogwerf, 2013). These numbers indicate a trend of changing views on sport club membership, and do not indicate a decline in interest in sports per se. However, it is interesting to research how sport clubs can respond to these numbers and changes in order to

lower dropout. It remains unclear however, next to the rough age group division, what other characteristics could be attributed to dropout. These reasons for dropout are personal reasons, still there might be reasons for dropout that are attributable to the social composition of the sport club, i.e. socioeconomic or demographic reasons.

The scientific relevance of this thesis is thus to establish a better understanding of sport dropout and to look into social composition aspects that could contribute to dropout. Sport marketing research already indicates that it is essential to focus on retaining members in order to survive as a sport club (Milne & McDonald, 1999). This becomes even more clear when looking at dropout numbers given by NOC*NSF: On a yearly basis about 130,000 people quit playing sports at a sport club according to the *Sportersmonitor 2012* (Hendriksen & Hoogwerf, 2013). However, estimations of the actual dropout size are inaccurate, as dropout in the *Sportersmonitor 2012* is measured by looking at the total difference in membership numbers from one year to another. For example, a sport loses about 10% of its members per year, but also gains that percentage every year, and thus shows no dropout. Still 10% of the members have left. The estimation of NOC*NSF only takes in account the 130,000 individuals that stopped playing sports, which is a net base called course in this thesis. The actual dropout of sport (club) members is much higher, which will be discussed for soccer later on. This failure to estimate dropout adequately and – in extension – to anticipate adequately on dropout also indicates that sport associations are more inclined to look for new members while it is less costly to satisfy the old members in such a way that they choose to stay (Milne & McDonald, 1999).

The societal relevance of this research is two fold. On the one hand, previous research only looks at the explicit individual reasons for dropout, most commonly the ‘don’t have time’-factor. This research will look into social composition of sport clubs to develop a better understanding of what the social composition effects on dropout entail. On the other hand, the role of sports in society has been researched a lot; sports in general play a crucial role in social cohesion, social capital of individuals as well as groups, and participating in sports is part of a healthy life style. Knowing more about what effect social composition of sport clubs has on dropout will also create a better understanding of how to create situations to lower dropout rates. Lowering dropout rates helps to enhance the societal benefits mentioned above (i.e. social cohesion, social capital, and a healthy life style).

This research looks into why people end their sport club membership. The aim is to look at the general factors that could explain this decision, related to the social composition of the sport club. Therefore this research tries to answer the following question:

To what extent does the social composition of a sport club play a role in the decision of members to end the membership?

The purpose of this research is to find out whether soccer clubs face higher dropout rates due to their social composition in terms of ethnicity, income, education level, gender or age. i.e. to understand the effects of these independent variables on dropout better. Therefore, the sub questions of this research are:

1. *What effect has ethnicity on dropout?*
2. *What effect has income on dropout?*
3. *What effect has education level on dropout?*
4. *What effect has gender on dropout?*
5. *What effect has age on dropout?*

To answer these sub questions this research first reviews what already has been reported on the effects of these independent variables. Subsequently, it looks into what is already in the data by reporting descriptive statistics. It answers the sub questions by doing statistical analyses (correlations and multiple regression).

In short, we do not know what the effects of social composition are on sport dropout. To create an initial understanding of the effects of social composition, this research uses data of the largest sport association in the Netherlands: the Royal Dutch Football Association (KNVB) and combines these data with datasets from Statistics Netherlands to answer the sub questions. Soccer was chosen as the key sport of interest based on the fact that it is a widespread practiced sport among all groups in Dutch society, and therefore was expected to show most variations in social composition of sport clubs. The KNVB has about 1.2 million members, which makes it the largest sports federation in the country, and plays an important role in the wider social context of the Netherlands (KNVB, 2014). The KNVB granted permission to use 7 years of membership data for the analyses, after discussing the exact

usage and signing a confidentiality agreement. Therefore, there are no direct references to soccer clubs or their members included in this study.

Outline

This thesis is built up out of four parts. Part 1 includes a conceptual clarification and a theoretical framework. The conceptual clarification explains what role sport has in society, discusses hunkering down, features of the sport club and membership, and explains social composition and dropout. The theoretical framework contains a discussion of the independent variables and forms the basis for the expectations. Part 2 includes the data and methods used to analyze the effect of the independent variables on dropout, and discusses the measurements used. Part 3 includes the analyses done: descriptive statistics, correlations and multiple regression. This part develops a model including the most important indicators that can be used to understand the effects of social composition on dropout. Bivariate and partial correlations are conducted, both normally and bootstrapped. Multiple regression is executed, creating a model of social composition to understand its effects on dropout. Part 4 shows a summary of the findings, a conclusion and discussion. This part also presents an encore, looking into possible directions for research on social composition using indexes based on heterogeneity and homogeneity.

2. Conceptual clarification

2.1 Sport & society

In the last couple of decades civic engagement has declined according to Putnam (2001). Civic engagement can be assessed by looking into voluntary associations. However, Putnam's decline of community thesis is not per se backed up in the Dutch context (Pharr et al., 2000; Schnabel, Bijl, & De Hart, 2008; Van Ingen, 2009). Later work of Putnam shows that he has found a way to conceptualize diversity without referring to a decline of community per se: "hunkering down" (see Putnam, 2007). He writes that in the long run immigration and diversity are likely to have important cultural, economic, fiscal, and developmental benefits. In the short run, however, immigration and ethnic diversity tend to reduce social solidarity and social capital. This means trust is lower, altruism and community cooperation become rarer, and friends are fewer. Albeit not all immigrants are ethnically different from the native population, in the Netherlands we use a distinction between Western and non-Western immigrants (see Statistics Netherlands, 2014b).

This research will look into the reasons for ending membership that are attributable to the social composition of a specific form of voluntary association: the sport club. As ending membership might be seen as proof of the decline of community thesis, looking into reasons for terminating the membership might reveal new information. Taking into account the social composition of the sport club will help to understand ending memberships (i.e. dropout). This research uses the theoretical perspective of hunkering down, rather than the decline of community thesis, as this is more appropriate in the Dutch setting.

Therefore, this research looks into voluntary but formal sport clubs, not taking into account other sport arrangements, such as fitness clubs, informal sportive activities (e.g. running together with a neighbor every now and then), or sport activities carried out alone in non-organizational settings (e.g. sportive cycling). For this research a number of concepts are important: hunkering down, features of the sport club, sport club membership, social composition, and dropout.

2.2 Hunkering down

According to Putnam diversity has a number of benefits. Firstly, creativity in general seems to be enhanced by immigration and diversity. Secondly, immigration is generally associated with more rapid economic growth. Thirdly, in advanced countries with aging populations, immigration is important to help offset the fiscal effects of the retirements of the baby-boom generation. Lastly, new research from the World Bank has highlighted yet another benefit of immigration: remittances.

There are two well-known theoretical perspectives on the influence of diversity on social capital. The first is the *contact hypothesis*, which entails that diversity fosters interethnic tolerance and social solidarity; if we have more contact with people of other ethnic and racial backgrounds, we will begin to trust one another more. This perspective thus suggests that diversity erodes in-group/out-group distinction and enhances out-group solidarity or bridging social capital, thus lowering ethnocentrism. The second is the *conflict theory*, which deals with contention over limited resources. This perspective suggests that diversity enhances the in-group/out-group distinction and strengthens in-group solidarity or bonding social capital, thus increasing ethnocentrism.

Putnam uses these theoretical perspectives to position himself in this discussion. He does this by using two different conceptions of social capital: bonding and bridging. Bonding social capital refers to ties to people who are like you in some important way. Bridging social capital refers to ties to people who are unlike you in some important way. Putnam points out that these two competing perspectives share the assumption that in-group trust and out-group trust are negatively correlated. Thus both contact hypothesis and conflict theory assume that bridging social capital and bonding social capital are inversely correlated in a kind of zero-sum relationship; if I have a lot of binding ties, I must have few bridging ties and vice versa. Putnam argues that this might not be true. Once we recognize that in-group and out-group attitudes need not be reciprocally related, but can vary independently, then we need to allow, logically at least, for the possibility that diversity might actually reduce both in-group and out-group - that is, bonding and bridging social capital, he calls this third possibility *constrict theory*. This constrict theory points out that increasingly social complex contexts can lead to hunkering down. While Putnam focuses on the effects of a diversification of ethnicity in society and its effects on in-group and out-group attitudes, this research also looks into other social variables that could influence those attitudes.

2.3 Features of the sport club

Associations or clubs are organizations formed around a shared goal that is related to a common interest in a topic, but not related to making a living. An association or club is used to structure free time for example, bridge clubs, women's associations, and so on.

Sport clubs are a special form of associations, where playing sports is the main reason for the association to exist. This could be any type of sport, but we are mostly talking about team sports and sports that are backed by a federation. Think about soccer, basketball, hockey, athletics, et cetera.

Sports clubs have different characteristics. In the first place there are sport clubs that provide team sports and (semi) individual sports. Team sports are sports that involve players working together towards a shared objective. Examples are hockey, soccer, rugby, basketball, handball, and water polo. Individual sports are sports in which individuals compete with each other to greater or lesser extent. Examples are athletics, badminton, boxing, taekwondo, and cycling. Both types of sports can be formally organized in a sports club.

Next to the team-individual division, sport clubs in itself can have other orientations towards individual characteristics. Membership recruitment is guided by these preferences, but is also affected by the financial structure present in Dutch local and national government (Stokvis, 1979). These different preferences of individual characteristics will be discussed in relation to sport participation and dropout in the theoretical framework. In addition, sport club membership is affected by socio-economic status, which is also reflected in types of sports. For example, soccer is played throughout all socio-economic status groups while hockey and tennis for example are deemed more elitist sports.

One of the reasons that certain people tend to engage in certain types of sports can be attributable to socialization processes. Bourdieu (1977, 1986) stresses the importance of institutions of socialization in systems of symbolic power. Looking at the Dutch sport club as an institution of socialization, we can expect that if socialization fails, people would dropout. Success factors of socialization are also dependent on contextual factors such as neighborhood characteristics. This also has implication when we expect that socialization in socially complex sport clubs is harder. One could imagine that a sport club with a lack of socialization has much more trouble in maintaining the sport club a desired level of membership.

This research focuses on soccer. The choice for soccer was made on the basis of the fact that it is the largest sport in the Netherlands, giving access to all regions, both urban and rural, and to all socioeconomic classes. In addition, the KNVB was willing to participate in this research and granted permission to use the membership information needed to conduct this study. The fact that soccer is a wide-spread sport, also has implications for the way socialization takes place. Stuij & Stokvis (2011) argue that high socioeconomic status groups have a socialization based on the nuclear family, whereas low socioeconomic status groups base their socialization on the extended family, creating very different dynamics.

2.4 Sport club membership

One of the features of a sports club is membership. There are a couple of people in charge of running the basic activities of the club (also administratively), and the rest are members. Clubs mostly run on volunteers that periodically change roles, for example, some people have a task on the board, while others are responsible for youth activities. Other members may only go to the club for sporting and meeting their teammates, while others help out at the bar every now and then.

Most sport clubs are focused on a specific age cohort. These cohorts could be defined in terms of youth (<18yrs), adults, and people over 45 years. About one-fifth of the sport clubs have a relatively large amount of adult members. The number of members is a concern for all sport clubs, even more so when focused on adults and the 45+ cohort (Tiessen-Raaphorst, Verbeek, De Haan, & Breedveld, 2010).

Membership of a sport club means at least paying the membership fee. Next to paying the fee, membership could be viewed as active membership, i.e. actually showing up at the sport club, engagement in a team and/or practicing on a regular basis, as well as taking part in tournaments or competitions. The first is easily derived from administrative efforts by the sport club, or its association. The latter is a bit more difficult to put into practice, especially when looking into ending memberships. This is because the termination of a membership might be preceded by low activity in the club. This research will only look into the first type of membership. In the case of soccer people become member of a soccer club and through the club they also become member a of the KNVB.

2.5 Social composition

The social contexts within which sports are practiced are an important indicator for the popularity of certain sports in certain social contexts while it may be of relative unimportance in others. Van Bottenburg (2001) explains that the social context of the sport is perhaps the most important factor in its rise (and fall) of popularity. He argues, in line with Stokvis' work, that there should be more focus on the characteristics of collectivities. Van Bottenburg (2001) has inspired this research in its focus on the relation between more macro sociological processes and individual sport preferences that are expressed in sport dropout. Building on his research, I look into a sport that is seemingly not influenced by class-linked preferences from a bird's-eye perspective. However, it might as well be that macro sociological influences can be deferred from "within." By this I mean that although the sport (soccer) is practiced amongst all social classes in the Dutch context, the characteristics of the sport club as a representative of the meso sociological level could have an impact on the rise (and fall) of popularity of that specific sport club amongst people that have common characteristics to a certain extent. Via this way, this research tries to link reasons for participation and dropout to social composition aspects of the sport club itself.

The social composition aspects are derived from the members of the sport club, to establish a base line per sport club. Therefore individual characteristics will play a central role in defining the social composition. There are a couple of sport club characteristics that need to be considered here. First, the size of the sport association has to be considered. Large sport clubs have different mechanisms than small sport club, especially when it comes to identification with the club as a whole. This might influence the decision to dropout for individuals. In addition, sports clubs can target certain age groups. This also influences the social composition, and the related activities to attract and keep members. Also, certain age groups are known to end their membership more than others (especially youth in the high school age end their membership). In addition, different age groups have different reasons for dropout. Overall, men are more often member of a sports club than women, so gender is also an aspect of social composition to be taken into account. Moreover, education level influences membership of sport clubs, with lower educated people partaking less often in sports than higher educated people. The same holds for income: people with a lower income face more difficulties participating in sports, while people with a high income participate in sports more often. Lastly, ethnicity is an important factor in the social composition of sport clubs, as

people with a non-western background are less often member than people with a western background. The social composition of the sport club could determine the level of dropout.

2.6 Dropout

To understand the meaning of dropout in the wider social context, I first look at another form of dropout: high school dropout. Dropout can be the outcome of alienation from school, representing a misfit between student needs and the school demand, and a deviance from expected development (Archambault, Janosz, Morizot, & Pagani, 2009). What is important in this type of dropout, is to understand the development of the student in a way that includes the facets of psychological experience, both for individuals and for groups (Archambault et al., 2009). Translating this to the sport dropout that is central in this thesis, alienation from the sport club could indicate future dropout, as well as the fact that both individual and group characteristics should be taken into account when analyzing dropout. As explained above, other authors have researched these individual aspects, and therefore the focus of this thesis is on group characteristics at the sport club level.

Kalmijn & Kraaykamp (2003) define dropout as leaving secondary school without a degree at the current level, while they point out that most scholars focus on premature dropout. As sports lack the obligatory aspect intrinsic in secondary schooling, dropout in sports is defined differently. In sports literature, the definition of Salmela is commonly used: “The term dropout implies voluntary premature dropping out of sport’s career i.e. sudden and unexpected quitting sport in a situation where an athlete did not use up entirely his/her potential” (Salmela, 1994 in Lepir, 2009, p. 194). This definition makes it possible to see what voluntary reasons for dropping out can be related to the sport club. However, Salmela’s definition is best applicable in professional sport contexts. Therefore, this research defines *dropout as voluntary premature dropping out of a sport club*, converting Salmela’s definition to be applicable in a recreational sport context.

3. Theoretical framework

In this theoretical framework I will discuss the different aspects of sport dropout in relation to social composition, including formulating expectations. To do so, I will first discuss how to understand this research in light of other research previously conducted.

An individual level perspective on sport participation shows that social support, economic status, gender, and life stage heavily influence the choice of an individual to partake in sports or to withdraw from sports (Lim et al., 2011). However, a system level perspective can help explain participation levels and dropout. Green et al. (2005) argue that sport development and participation patterns can be impacted by the design and implementation of the sport delivery system itself.

The sport delivery system in the Netherlands has been provided outside the schools by private but voluntary sport clubs, and these clubs predominantly served a young, middle-class, white, male population (Lim et al., 2011; see also: Van Bottenburg 2001). Lim et al. (2011) found that data from the Netherlands indicates that overall sport participation steadily declines with age (see also Van Bottenburg, Rijnen, & Van Sterkenburg, 2005), however in terms of gender Dutch women and men participate in sports at a comparable level. Sport clubs are typically driven by memberships, providing sport opportunities for both youth and adults (Elling, Knoppers, & Knop, 2001). Sport in the Netherlands has become a normal leisure time activity, as about two-thirds of people over the age of 16 participate in some form of sports weekly (Van Bottenburg, 2001). One reason that sport participation is rather high in the Netherlands is because of the sport club system, which provides easy and lifelong access to sports (Lim et al., 2011). The Royal Dutch Football Association could be seen as such a sport delivery system.

This research taps into these ideas on individual level and system delivery level, but situating itself in the middle, with influences from both individual level factors and system level factors: the sport club. The sport club will be researched using individual level characteristics, that combined define the social composition of the sport club, to assess the relation between social composition and dropout.

3.1 Sport dropout & ethnicity

There is relatively little known about the influence of ethnic background on sport participation and dropout. The Netherlands Institute for Social Research (*Sociaal Cultureel Planbureau*) states that sport participation is unevenly distributed amongst ethnic groups in the Netherlands. Participation is highest among *autochtonen*, and lowest among Turkish and Moroccans, which is also reflected in sport memberships (Schnabel et al., 2008). Interethnic contacts are enhanced by sport activities among all ethnic groups. In addition, the *Verenigingsmonitor 2008* (Kalmthout, Jong, & Lucassen, 2009) has found that more than half of the sport clubs have *allochtone* members. Medium-sized and large sport clubs that offer team and semi-individual sports have more *allochtone* members than small associations and associations that offer individual sports. There are relatively more sport clubs with *allochtone* members in larger municipalities and clubs located in the western part of the country (Kalmthout et al., 2009).

The voluntary nature of sport clubs is linked to social capital (see Putnam, 2000), which is widely discussed in studies of ethnic sport participation as social capital leads to useful contacts, knowledge, skills, and trust (Verweel, Janssens, & Roques, 2005). Verweel, Janssens & Roque (2005) in their study of *autochtone*, *allochtone* and mixed sport clubs, point out that sport clubs are a context in which social networks take form. Social capital gains that are made within the sport club context are also useful outside that context. If sport participation is related to social capital gains, then sport dropout might be related to the absence of these kind of gains. However, social capital might not be the primary reason for different ethnic groups to participate in sport club, or to end their membership. Still, the effects of ethnicity on sport participation could also inform us about how to understand sport dropout from an ethnicity perspective.

In the contextual clarification the effects of interethnic contact have been discussed, and the term ‘hunkering down’ was coined. In this research the expectations on the effects of ethnicity on sport dropout are:

E1.1: High percentages of Westerners lead to low dropout rates

E1.2: High percentages of non-Westerners lead to high dropout rates

3.2 Sport dropout & income

As previous research shows, lack of money is one of the main reasons to quit playing sports (Tiessen-Raaphorst et al., 2010). However, very little is known about the relation between income and sport dropout. *Rapportage Sport 2010* (Tiessen-Raaphorst et al., 2010) shows that the stability of a fixed income creates possibilities for having children for example (and thus perhaps less time to do sports). People with a low income participate significantly less in sports than people with a high income (Tiessen-Raaphorst et al., 2010).

Jehoel-Gijsbers (2004) explains that people with an income lower than 130% of the social minimum have significantly more chance not to participate in sports. Extending this finding to sport dropout, could indicate that changes in income, especially towards 130% or less of the social minimum income can cause dropout. In the contextual clarification the effects of diversification in groups was discussed, causing hunkering down – an important factor to withdraw from sports because of factors such as lack of identification, recognition, and clarity of expected behavior. As sport dropout could be driven by financial means, the expectations for income are:

E2.1: High minimum income leads to low dropout rates

E2.2: High average income leads to low dropout rates

E2.3: High maximum income leads to low dropout rates

This independent variable is controlled for ethnicity as well, as there might be an overlap between ethnicity and income in their explanatory value for dropout.

3.3 Sport dropout & education level

Tiessen- Raaphorst et al. (2010) state that sport participation is influenced by education level. This is shown in the observation that higher educated people are sportspersons than are lower educated. Higher educated groups have around 60% sport participation only declining after retirement age, whereas lower educated groups have between 30-40% sport participation. Also, membership numbers will be relatively uninfluenced by demographic developments such as ageing since the education level has increased (Tiessen-Raaphorst et al., 2010). Sport participation in itself tells nothing about sport dropout, however, when looking at education

level, interesting differences in sport participation can be observed throughout the life course. In the conceptual framework the term ‘hunkering down’ is discussed as an effect of the uncertainty of behavioral expectations among different groups of people. One could argue that educational background influences behavior in a similar way, causing clear expectations in-group and unclear behavioral expectations out-group. Linking this to what is expected from different education levels, this research expects that:

E3.1: High percentages of low educated people lead to high dropout rates

E3.2a: High percentages of medium educated people lead to high dropout rates

E3.2b: High percentages of medium educated people lead to low dropout rates

E3.3: High percentages of high educated people lead to low dropout rates

This independent variable is controlled for ethnicity, as there might be an overlap between ethnicity and education in their explanatory value for dropout.

3.4 Sport dropout & gender

Sport participation and dropout can be related to gender differences and stereotyping. Boiche, Plaza, Chalabaev, Guillet-Descas, & Sarrazin (2013, p. 1) state that “[g]ender differences in sport are often perceived as resulting from natural biological factors.” Nevertheless, the article shows that gender differences in sport can be traced back to social processes. Existing literature has observed gender difference in levels of perceived competence and value in sport across age and culture. Gender stereotypes in sports are likely to have an impact on self-perception and behavior in sport activities (Boiche et al., 2013), which ultimately can lead to dropout. As Boiche et al. (2013) state: “Indeed, it appeared that adopting a gender-biased view of sport could significantly predict intentions to dropout from sport, through indirect effects of self-perceptions (regarding competence and attainment value) in the sport context” (2013, p. 13). However, not only stereotyping is important for understanding gender aspects of sport dropout.

Scanlan, Russell, Magyar, & Scanlan (2009) show that sport commitment (or lack of it) is an important factor as well. Sport enjoyment strengthens commitment without gender differences, and is the most influential factor for commitment (Scanlan et al., 2009).

Soccer used to be an all-male sport, however in recent years the number of female soccer club members is on the rise. Due to these gender differences, most female soccer members are expected to be relatively young. However, the change of soccer into a mixed sport could have advantages for the membership level, especially at critical life cycle moments such as changes of school, puberty, and finding a partner. In this research the expectations of the effects of gender on dropout are:

E4.1: High percentages of males lead to low dropout rates

E4.2: High percentages of males lead to high dropout rates

E4.3: High percentages of females lead to low dropout rates

E4.4: High percentages of females lead to high dropout rates

These expectations need to be controlled for age to see particular changes in dropout.

3.5 Sport dropout & age

In different age groups motivations for sport participation are different. In the *Rapportage Sport 2010* (Tiessen-Raaphorst et al., 2010) sport participation across the life course is discussed. In the first years of life, sport participation is pretty much influenced by parents. In the years a child frequents elementary school, sport participation is still influenced by parents, but also affected by friends. Between the ages 9 to 12 a child develops its own feeling for sports, which sports he or she is good at and which ones he or she (dis)likes. During the teenage phase, sport participation is influenced by identity development, peer groups and sexuality. In the cohort 18 to 34 year-olds education level and ethnicity play an important role. People with a higher education and Dutch background participate in sports more often. In addition, male sport participation is for this cohort significantly higher than women's participation. In the cohort 35 to 64 year-olds women participate more in sports than men. Also, people who are employed or are busy with the household participate more in sports than people who are unemployed or incapacitated. Education level and household income are also determinate factors for higher sport participation. Sport participation in the cohort 65+ is determined by their physical abilities. Also, partnership has a positive influence on doing sports. Lower household income has a negative correlation with sport participation. In

general, adults who do not do sports are more often of a non-western background, are obese, and have a lower education level (Tiessen-Raaphorst et al., 2010).

A large number of scholars have researched sport dropout in adolescence, which is logical since this age cohort is dropping out a high rate. It also points to a major knowledge gap when it comes to sport dropout in other age cohorts. This is noteworthy as participation rates steadily decrease as people age (Gucciardi & Jackson, 2013). Fraser-Thomas et al. (2008) report that for adolescents the most commonly cited reasons for dropout are conflicts of interest, negative experiences such as lack of fun, coach conflicts and lack of time. The latter in combination with lack of interest is most prominent for sport dropout according to their study. Another study carried out by Boiche & Sarrazin (2009) indicates proximal and distal predictors of sport dropout. Proximal factors are the most important reasons for dropout such as perceived value of the activity, satisfaction and parents' investment. According to their study the duration of sport participation is influenced by "(1) demographic or biological characteristics (e.g., sex, age, BMI), (2) psychological or cognitive attributes (e.g., motivation, perceived competence, intentions of participation), (3) social and cultural factors (e.g., social support) and/or (4) environmental contingencies (e.g., opportunities to exercise, equipment available)" (2009, p. 9).

In general there are different age cohort distinguishable in adult life: young adults are still studying (18-24 years) or focusing on their careers (25-34 years). After this period most adults try reconcile work and family life with small children (35-44 years) or older children (45-64 years) (Tiessen-Raaphorst et al., 2010). Due to these differences in main occupation during the life course, it can be expected that reasons for dropout are also different for the different age cohorts. Especially the emergence of informal sport groups makes it easier to end formal memberships of a sport club (Tiessen-Raaphorst et al., 2010). The discussion on sport club participation and possible reasons for dropout in different age cohorts has not lead to a theoretical base for formulating directional hypotheses. Therefore, this research expects that:

E5.1a: High minimum age leads to low dropout rates

E5.1b: High minimum age leads to high dropout rates

E5.2a: High average age leads to low dropout rates

E5.2b: High average age leads to high dropout rates

E5.3a: High maximum age leads to low dropout rates

E5.3b: High maximum age leads to high dropout rates

3.6 Theoretical model

These expectations can be summarized in the following model:

Independent (high values of)	→	Outcome
Western ethnicity	-	Dropout
Non-Western ethnicity	+	
Minimum income	-	
Average income	-	
Maximum income	-	
Low education	+	
Medium education	-	
Medium education	+	
High education	-	
Male	-	
Male	+	
Female	-	
Female	+	
Minimum age	-	
Minimum age	+	
Average age	-	
Average age	+	
High age	-	
High age	+	

4. Data and Methods

4.1 Data description

4.1.1 Dataset Statistics Netherlands – Postal codes and ethnicity

This dataset consists out of the postal code-format used in the Netherlands with six characters (1234AA, called PC6) and ethnicity. Ethnicity is measured by rounded percentages of non-Western inhabitants (“*allochtonen*”) compared to the complete population, measured on January 1, 2004. This data is retrieved from the municipal administration. *Allochtonen* are divided into Western and non-Western on the ground of their country of birth. The category non-Western include people from Turkey, Africa, Latin America, and Asia (excluding Indonesia and Japan, and the Asian countries that used to be part of the Soviet Union). *Allochtonen* who are born in a foreign country are called first generation. The second generation is determined by the country of origin of the mother. If this is the Netherlands, the country of origin of the father determines the ethnicity status. A person born in a foreign country but with Dutch parents is considered Dutch. Percentages are determined with a minimum of 10 inhabitants in the postal code area and shown in different classifications. Classification 1: less than 5% non-Western; classification 2: 5-10% non-Western; classification 3: 10-20% non-Western; classification 4: 20-40% non-Western; classification 5: more than 40% non-Western (Statistics Netherlands, 2014b). This dataset will help to determine the balance between Western and non-Western members for the postal code of the soccer club and for the average of all postal codes of the members of the soccer club.

4.1.2 Dataset Statistics Netherlands – Postal codes and income

This dataset consists out of the postal code-format used in the Netherlands with six characters (1234AA, called PC6) and income before taxes. Income before taxes (“*fiscaal maandinkomen*”) is calculated by income from work, assistance, and pensions of a person at the end of the year 2008. Of these sources of income the yearly wages and their payment

periods are known. From these sources the averages are calculated, and the different sources are cumulated. The data is retrieved from data known by the Dutch tax office. The averages are shown for all postal codes with more than ten (10) income recipients in a particular postal code. The averages are rounded to hundreds with a minimum of €500 per month and a maximum of €10,000 per month (Statistics Netherlands, 2014b). This dataset will help to determine the average income per soccer club, both by looking at the averages of the postal code of the soccer club and by looking at the averages of the collective postal codes of their members.

4.1.3 Dataset Statistics Netherlands – Postal codes and education level

This dataset consists out of the postal code-format used in the Netherlands with four digits (1234) and education level (Statistics Netherlands, 2014a). Education level is measured by the highest education with a diploma on September 30, 2011. Education level is divided into three categories: low, medium, and high. Low education level means primary education and lower secondary education, including VMBO (pre-vocational education), lower classes of HAVO (higher general secondary education), and first three classes of VWO (pre-university education). Medium education level means the higher classes of secondary education (HAVO and VWO), and MBO2, MBO3, and MBO4 (vocational education). High education level means HBO (university of applied sciences) and WO (university). Areas with a lot of young people, especially those under 16 years, could be influenced by their low education level. This dataset will help to determine the education level distribution per soccer club, both by looking at the postal code of the soccer club and by looking at the averages of the collective postal codes of their members.

4.1.4 Dataset Royal Dutch Football Association (KNVB) – Member information

This dataset includes data of all members of the KNVB in the period 2006-2014. In the beginning there were eight different datasets, one for every season. Every season held the same types of information: relation code, postal code, date of birth, gender, status, mode, club, season, and reference date. These eight datasets were combined into one dataset (*vereniging per seizoen*). Relation code is a unique number given to a particular member of KNVB. This number does not change if a person is member for consecutive years, including

transfers. However, the relation code does change if a member unsubscribes for a year, and becomes a member again later. Tracing the relation code over a period of time helps to determine course (*verloop*) and dropout. The postal code is known in the format 1234AA (PC6). The postal code is related to the datasets of Statistics Netherlands in order to be able to say something on income, education, and ethnicity influences on dropout. The date of birth is given in the YYYY-MM-DD format, and is used to calculate age. Age is important to include in the analysis in order to control for age-specific differences. Gender can be male, female or unknown (unknown is statistically negligible). Gender is important to include in the analysis in order to control for gender-specific differences. Status can be either playing or non-playing, and refers to the members who are active (playing) at a soccer club, or those who are non-active, e.g. volunteers, but also those people who do not regularly visit their club. Mode means the type of soccer played by the member and can be roughly divided into field or hall. Field can be divided into Saturday and Sunday. Season refers to the season of the reference date. The reference date is set on April 30, for the corresponding season. This dataset gives information relevant for an analysis of club specific configurations and their influence on individual dropout and overall dropout levels of the Royal Dutch Football Association. This data was made available by the KNVB.

4.1.5 Dataset Royal Dutch Football Association (KNVB) – Club information

This dataset includes data of all clubs registered with the KNVB in the period 2006-2014. Every club holds the same kind of information: club code, district, name, founding date, end date, and information on the number of Saturday, Sunday and hall soccer per season (period 2006-2014). This dataset gives information relevant for an analysis of the club environment, and to be able to compare this environment with the averages of the members of a particular club. This data was made available by the KNVB.

4.1.6 Dataset Royal Dutch Football Association (KNVB) – Club postal codes

This dataset includes data of all clubs registered with the KNVB in the period 2006-2014. For every club their postal codes in 1234AA format is available, their club code, and their name. This dataset is necessary to be able to link datasets, as will be explained below. This data was made available by the KNVB.

4.2 Methodology

4.2.1 Research Intelligence

The datasets were included in a program for business intelligence in order to create a sophisticated model that included all the different variables of interest: QlikView.

QlikView is a business intelligence software program that helps to visualize large pieces of data. QlikView is also associative in its way of working, meaning that it creates associations within and between different types of data inserted in the program. In addition, QlikView is reactive as it selects those pieces of information you click on. In this research QlikView is chosen to visualize and analyze data as different datasets are used, and a visualization of the data helps to understand the possibilities of the data better (see descriptive statistics below). In addition, the KNVB dataset was not a sample, but information on the complete population which makes statistical analysis on the level of one season unduly.

Datasets postal codes and income, postal codes and education, and postal codes and ethnicity, were scanned twice, once to link the information to the individual members and once to link the information to the clubs. (For a complete image of the QlikView model, see appendix 1.)

The core of the model is based on the dataset of the KNVB about member information (*Ledenonderzoek*). This dataset is linked through the club code (*Verenigingscode*) with the dataset on clubs (*VerenigingPerSeizoen*) and the dataset that contains the postal code information of the clubs (*PC6verenigingLink*). In addition, the postal code of the club (*PCVereniging*) connects the datasets *Inkomen_Vereniging*, *Demografische_Vereniging*, *Opleiding_Vereniging* and *Postcode_Vereniging*, which provide information about the members on the respective topics of income, ethnicity, education level, and postal codes. The same type of datasets were created for members and those are linked by the postal code of the member (*PC6Lid*). This connects the datasets *Inkomen_Lid*, *Demografische_Lid*, *Opleiding_Lid* and *Postcode_Lid*, which provide information about the members on the respective topics of income, ethnicity, education level and postal codes. The independent and dependent variables were requested in a table per club, and exported to SPSS.

4.2.2 Statistical analyses

Correlations

The correlation method is used to determine the extent to which the independent variables are related among a dropout scores, however there is no attempt to manipulate the variables. Thus, correlation research asks the question: what relationship exists? There are two main points here: Firstly, a correlation has direction and can either be positive or negative. A positive score indicates that a score on the independent variable scores similarly on dropout. A negative score indicates that a score on the independent variable scores oppositely on dropout (Siegle, 2014). Secondly, a correlation can differ in the degree or strength of the relationship. Zero indicates no relationship between the two measures and $r = 1.00$ or $r = -1.00$ indicates a perfect relationship. The strength can be anywhere between 0 and 1.00. As a rule of thumb I use the following guidelines for assessing r :

Value of r	Strength of the relationship
-1.0 to -0.5 or 0.5 to 1.0	Strong
-0.5 to -0.3 or 0.3 to 0.5	Moderate
-0.3 to -0.1 or 0.1 to 0.3	Weak
-0.1 to 0.1	None or very weak

Table 1. Guidelines for assessing r , Explorable (2009b).

A few things to keep in mind with regard to correlation coefficients (r): firstly, correlation coefficients mostly only show linear relationships. Secondly, correlation coefficients do not have to make sense to achieve an acceptable value. Thirdly, correlations only describe the relationship; they do not prove cause and effect. Correlation is a necessary, but not a sufficient condition for determining causality (Explorable, 2009b).

In this case, the correlation investigates the question “what is the relationship between social composition of a sport club and its dropout rate?” This question is partially answered through correlations, and expressed in bivariate and partial correlations. In order to overcome assumptions of the data, bivariate correlations expressed in Spearman’s Rho and partial correlations expressed in Pearson’s r using the bootstrapping method.

Multiple regression

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable; it is a techniques used for predicting the unknown value of a variable from the known value of two or more variable (Sawasthi, 2000). In this research the independent (predictor or exploratory) variables are ethnicity (Western, non-Western), income (average, lowest, and highest), education (low, medium, and high), gender (male, female), and age (average, lowest, and highest). All variables, including the dependent variable dropout, are expressed at the ratio level (either in percentages or in absolute numbers).

The regression analysis will show a value for b . b_0 is the intercept and b_1, b_2, b_k are analogous to the slope in linear regression, called regression coefficients (Explorable, 2009a). The appropriateness of the multiple regression model as a whole can be tested by the F-test; a significant F indicates a linear relationship between the dependent variable (Y) and at least one of the predictors (X). The predictive ability of the regression model is assessed by examining the coefficients of determinations (R^2); the closer R^2 is to 1, the better the model and its prediction. Multiple regression also shows if the independent variables (predictors) individually influence the dependent variable significantly (while controlling for the other variables in the model) using the t-test. If the t-test of a regression coefficient is significant, it indicates that the variable in question influences Y significantly (Explorable, 2009a). However, multiple regression in itself does not test whether data are linear, this is thus assessed separately, along with other assumptions (no multicollinearity, heteroscedasticity, and normality). With multiple regression I hope to answer the question “what is/are the best predictor(s) of dropout?”

Factor analysis

Social science researchers often try to measure things that cannot be measured directly (Field, 2009), factor analysis can help overcome this problem. In this research, factor analysis is used to find latent variables that might be in the data. SPSS is ordered to perform a principal components analysis (PCA) and all variable groups with Eigenvalues over 1 could imply a latent variable. This technique helps to understand the underlying mechanisms of social composition better; it answers the question “what predictors are latent variables (if any)?”

4.2.3 Methodological reflection

Normal science is a cumulative enterprise (Kuhn, 1962), but the current research, although related to other aspects of sport research – especially sport participation and adolescent sport dropout – broaches a new topic: non-individual related aspects of a sport club that could contribute to the decision of an individual to dropout. This is also an aspect that Flyvbjerg (2001) points out when talking about how to make social science matter: it is about letting go of the natural sciences ideas of cumulative and predictive theories.

Yet, previous research I conducted – schooled in cultural anthropology – always deviated strongly from a natural science approach. Therefore, developing a social science perspective that is based on more exact and less interpretative methods has been quite the challenge for me. The current research has made me aware of the fact that even seemingly straightforward, one-outcome methods such as correlations and multiple regression establish interpretable models, which are to be explained by the researcher. I see the merit in trying to find ways of explaining the phenomena that social scientists research by creating models.

Models in the actual science are neither derived from data, nor from theory, better models are understood as preliminary theories in my opinion. As, once we have knowledge of the model, this knowledge can be translated to knowledge about the social reality. During the creation of the model, its representational function is of less importance, but after the model is established its representational function becomes important again (see Plato Stanford, 2014). Does the model fit in any way to the social reality the social scientist liked to understand better? Initially creating a model was not the goal of this research, and therefore the end product is not a model. However, in the process of doing statistical analyses, a model occurred. This model can now be viewed as a preliminary theory and can be explored further by other social scientists that are interested in the relation between (sport) dropout and social composition of groups. Creating such a model can advance the understanding of social composition on dropout, even outside the researched context, such as different sports, different voluntary organizations, and so on.

As a final note of this methodological reflection I would like to point out the importance of the connection between science and practice. Following Flyvbjerg (2001), I hope to create scientific knowledge that is of practical use, paving the road towards a social science that matters.

4.3 Operationalization & measurements

The unit of analysis for this research is the sport club, and not the individual members. The outcomes of the analyses will express what social composition characteristics of soccer clubs contribute to dropout, but this does not indicate that people with a certain social composition characteristic are also the one who dropped out.

4.3.1 Ethnicity

Ethnicity is derived from a dataset made available by Statistics Netherlands, and contains information on two ethnicity types: Westerners, and non-Westerners. For these two ethnicity types the averages are calculated on a national level, on the aggregate club level, and on the level of specific clubs. The specific club level ethnicity is expressed as a percentage of Westerners and non-Westerners.

4.3.2 Income

Income is derived from a dataset made available by Statistics Netherlands, and contains information on income rounded to hundreds. The national income is calculated by all PC6 that occur in the KNVB datasets, and is settled at €2,462.09 (see Statistics Netherlands, 2014b). The club income is calculated by the income of their members using their respective PC6 data, and is expressed in absolute numbers (Euros). In addition, the income distribution is calculated per club by showing lowest income and highest income found with their members using the PC6 of the member.

4.3.3 Education level

Education level is derived from a dataset made available by Statistics Netherlands, and contains information on three education levels: low, medium, and high. For these three education levels the averages are calculated on a national level, using all clubs, and on club level. The club level education level is expressed as a percentage.

4.3.4 Gender

Gender is derived from the membership data of the KNVB. Gender is measured by the ratio of men and women per club compared to the national average, and expressed as a percentage. The national percentages are 90.14% for men, and 9.84% for women. The clubs are compared to the national average by taking the club percentage of male and female (dropout) members.

4.3.5 Age

Age is derived from the membership data of the KNVB. Age is calculated by taking the average age per club. The club average age is calculated by the age of their members, and is expressed in years. In addition, the age distribution is calculated per club, showing lowest age and highest age.

4.3.6 Dropout

Dropout is measured by looking at what people are not a member of a particular sport club anymore in the following year, e.g. someone who has been a member of a sport club in season 2006-2007, but is no longer a member on season 2007-2008. Dropout can be divided into people who stopped playing soccer at a club or transferred to another (dropout *Vereniging*), those people who stopped playing soccer at a club altogether (dropout KNVB), and those people who stopped playing at a certain club and transferred to another club in consecutive years, so-called transfers (dropout transfer). The first kind is the dropout that is used in this research, as the level of measurement is soccer clubs. However, people who stopped playing at the KNVB are discussed as they show us how many people dropout of soccer as a sport. In addition, transfers are discussed as this group is usually considered dropout at the club level, but form a special case when it comes to playing soccer at a club in general. Dropout rates are established as a percentage of the total membership per club.

5. Descriptive statistics

Before analyzing the data using SPSS, some descriptive statistics will help to establish benchmarks. These benchmarks contribute to assess the data analyses on the independent variables: ethnicity, income, education level, gender, and age, as well as the dependent variable dropout. In this section, these independent variables will be discussed by national averages if applicable, and by KNVB member and/or club averages. In addition, dropout will be discussed by looking at membership in general and at dropout in absolute and relative terms.

5.1 Ethnicity

In terms of ethnicity, sport participation has been researched extensively. Ethnic minorities in the Netherlands are participating less in sports, than people from Dutch descent. However, looking at the postal codes, this difference is not visible in soccer. The national averages in terms of ethnicity are 89% Western and 11% non-Western, while soccer club members have the same division (figures 1 and 2). Still, postal codes suggest that less non-Western people are playing soccer when looking at the division between Western and non-Western at the club level, which is respectively 94% and 6% (figure 3). In this research, ethnicity is used to establish the social composition of the sport club, which is informed by the ethnicity of the members linked to their individual postal codes. This means that there are soccer clubs with higher and lower percentages of non-Western members than 11%.

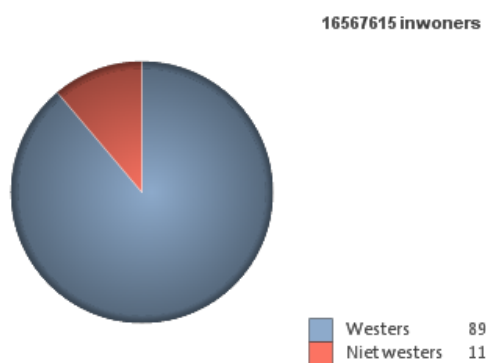


Figure 1. Ethnicity national figures

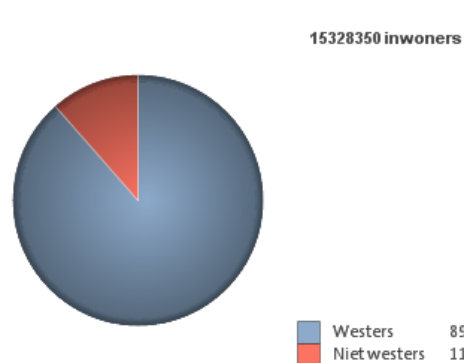


Figure 2. Ethnicity KNVB member figures 2006-2013

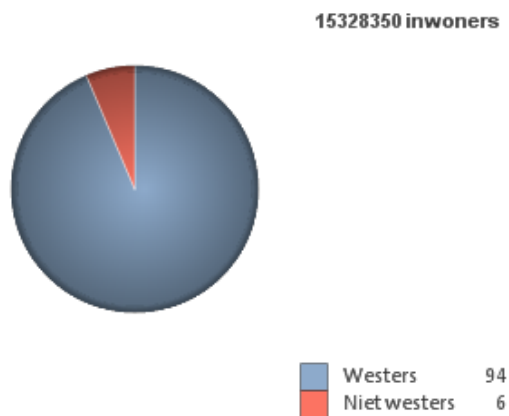


Figure 3. Ethnicity KNVN soccer club figures 2006-2013

5.2 Income

The income of an individual says something about how much they have to spend. Deviations from national and KNVB membership averages also tell us something about the type of people in a sport club, i.e. do they have more or less money to spend on average than the national income. The national average income derived from the data is €2,462.09, displayed in table 1. The KNVB membership average income is €2,486.61 per member (see table 2), slightly higher than the national average income. These averages will serve as benchmarks to assess the distribution of income within sport clubs, i.e. the lowest and highest income found within one club.

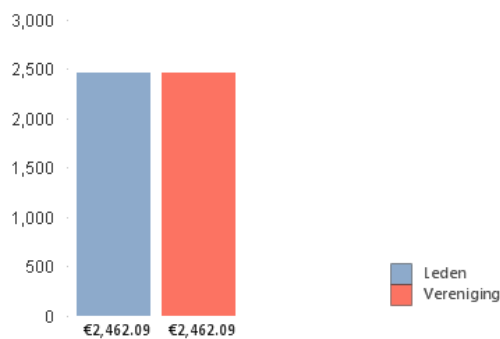


Table 1. Average national income

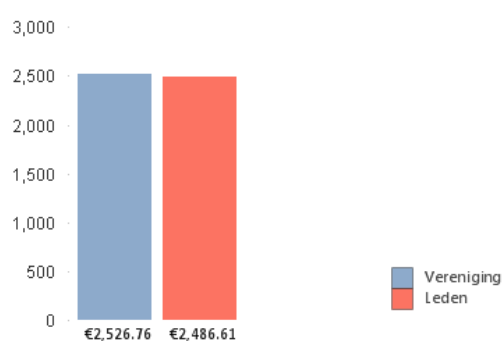


Table 2. Average income per club and member 2006-2013

5.3 Education level

Unfortunately, education was only accessible at a PC4-level, which means that the information available is less detailed than the other Statistic Netherlands datasets, as a larger area is covered per datum. Still, we can say something about the composition in terms of education level in the Netherlands, of soccer as a whole, and, later on, at the level of the sport club. The education level stays the same for all seasons at the national level, as there is only one point of measurement for this dataset and the PC4s included do not change the percentages overtime. The national distribution of education level is 47% low, 35% medium, and 18% high (table 3). Note that these numbers are derived from the population as a whole, thus influencing the scores as minors have a low education level per definition. If this were not the case and only adults (25-65 years) would be selected, these the numbers would be low education 27%, medium education 40%, high education 32% (Den Hertog, Verweij, Mulder, Sanderse, & Van der Lucht, 2014).

On average, 48% of the members of the KNVB have a low education level, 35% a medium education level, and 17% has a high education level (table 4), which resembles the national distribution. Again, these numbers are based on the education level of the soccer club member even in case of a minor, as information on the education level of the parents is unknown.

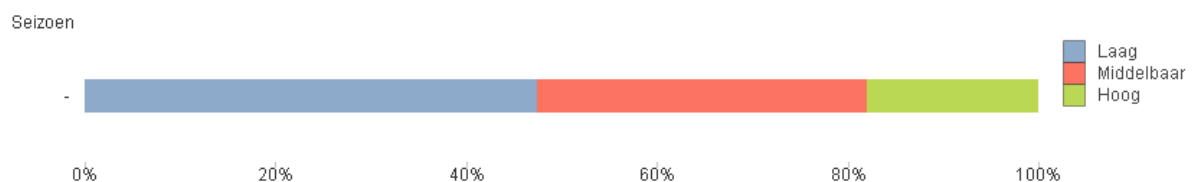


Table 3. Distribution of education level in the Netherlands

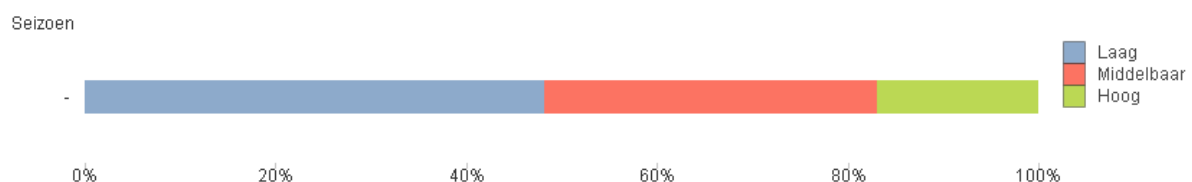


Table 4. Distribution of education level for all clubs for all seasons

5.4 Gender

Soccer is widely known as a male sport, however in the past years more and more females are member of a soccer club. This trend is also clearly visible in the data provided by the KNVB.

Between seasons 2006-2007 and 2013-2014 the percentage of women participating in soccer clubs rose almost 2.5% from 8.48% to 10.97%. In absolute

numbers the amount of women rose from 95,805 in 2006-2007 to 130,296 in 2013-2014, which is an increase of 34,491 women. The dataset provided also had a small amount of people of whom the gender was unknown or different from the male/female division, however this part of the population was insignificant ($p > .05$). Also, the amount of people of which the gender was unknown is reduced to none in the last observation season (2013-2014).

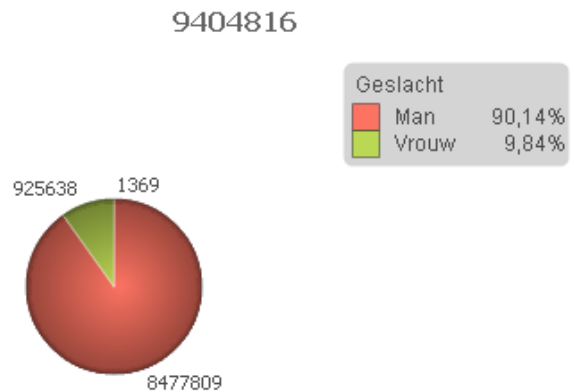


Figure 4. Gender pie chart 2006-2013

	Male	Female	Unknown
2006-2014	90.14%	9.84%	0.02%
2006	91.50%	8.48%	0.02%
2007	90.94%	9.04%	0.02%
2008	90.53%	9.45%	0.02%
2009	90.16%	9.82%	0.02%
2010	89.88%	10.10%	0.02%
2011	89.75%	10.24%	0.01%
2012	89.45%	10.54%	0.01%
2013	89.03%	10.97%	0.00%

Table 5. Gender percentages per season

One explanation for the rise in the level of female soccer club members can be found in the shift in gender stereotypes revolving around playing soccer. In the last years, female soccer participation has become more common, and thus it can be expected that male-female ratios are shifting even more in future years. However, a more detailed analysis of gender and their possible factors of influence should be researched in order to formulate more concrete outcomes.

5.5 Age

Age is calculated by the date of birth, and is related to the latest observation date (April 30, 2014). On average in the period 2006-2013 soccer is played mostly by youth, starting at the age of 5 and peaking at the age of 14, then rapidly declining until the age of 18, and then slowly declining with a small peak at the age of 45. This age distribution (table 6) shows that the most dropout would be located between the ages of 15 and 18 years. However, since there is a small increase in membership between the ages of 35 and 45, there is also a higher dropout rate after 45. These two critical moments (15-18 years and >45) could be further investigated in order to create policies to keep people of these ages more engaged.

The age distributions for particular seasons look quite similar, however, one can retrieve that there is stabilization around the mid-twenties, and in the latest years even a small increase. This aspect could be further investigated in order to understand this prolonged (and increased) membership better, i.e. investigating cohort effects.

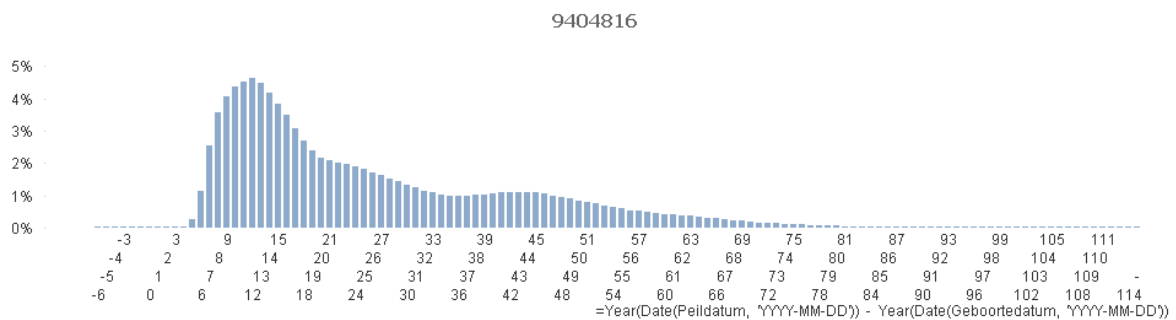


Table 6. Histogram age distribution 2006-2013

5.6 Members & dropout

To understand dropout better general membership of the KNVB is discussed. In the period 2006-2014 the membership has increased from 1129178 to 1188245, an increase of 4.97% (table 7 and 8). The growth in membership has been steadily rising in the period 2006-2012, but has decreased slightly in the season 2013-2014. This research tries to understand this breakpoint on the level of age, gender, ethnicity, and income of individual members in relation to the social composition on the same characteristics. The graph and table below show that the steady increase in members is mostly due to an increase of female members.

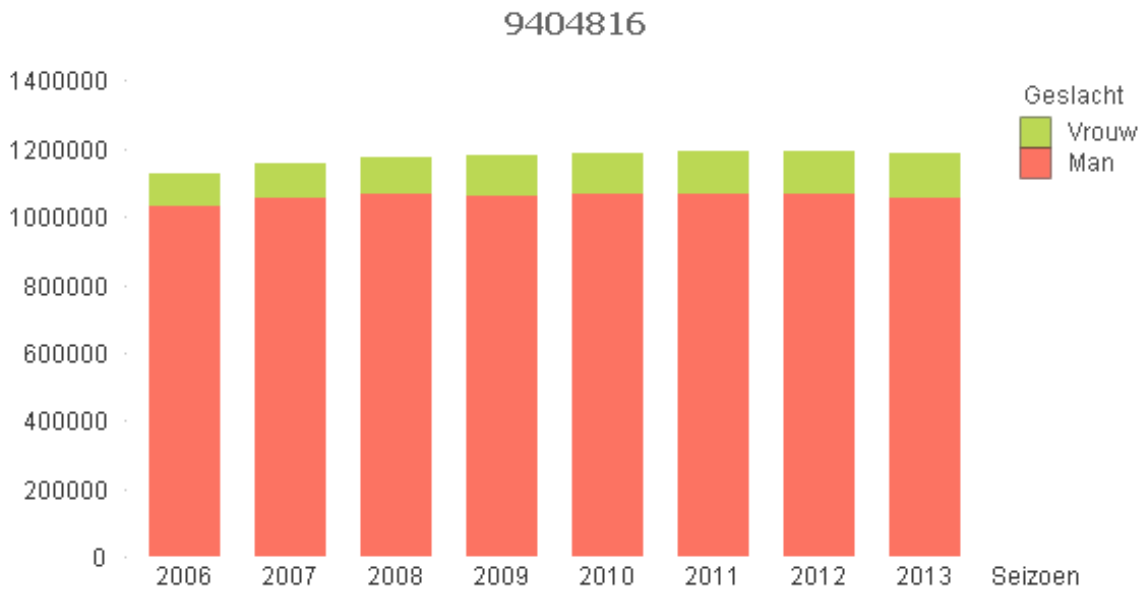


Table 7. Members by gender 2006-2014

Season	Male	Female	Total
2006-2007	1,033,373	95,805	1,129,178
2007-2008	1,053,480	104,770	1,158,250
2008-2009	1,066,567	111,326	1,177,893
2009-2010	1,062,750	115,754	1,178,504
2010-2011	1,066,790	119,938	1,186,728
2011-2012	1,068,327	121,888	1,190,215
2012-2013	1,068,209	125,861	1,194,070
2013-2014	1,057,949	130,296	1,188,245

Table 8. Members by gender in absolute numbers

Table 9 shows the course of membership for the period 2006-2013. This table indicates that KNVB membership (through clubs) rose every season, except season 2013-2014. The numbers indicate the change in memberships for every season compared to the former season.

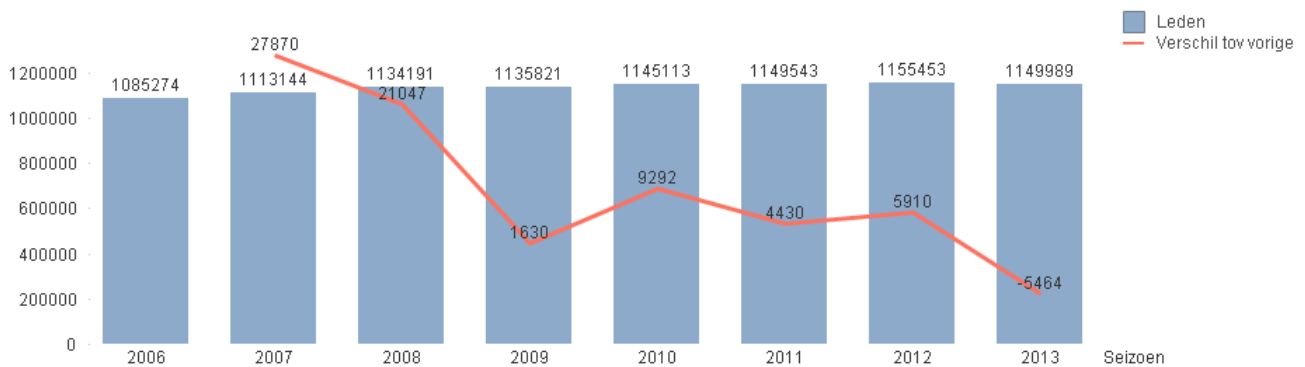


Table 9. Membership course KNVB 2006-2014

5.6.1 Dropout

Dropout at the club (*vereniging*) and KNVB level increases over the years in absolute (table 10a and 10b) and in relative numbers (table 11a and 11b). Dropout *vereniging* increases from 13.48% in 2006 to 18.15% in 2012, dropout for season 2013-2014 is unknown as the data for the next season is not yet available. Dropout KNVB increases from 8.32% to 12.88% in the same period. However, the transfer rate stays relatively unchanged, around 5%.

This research focuses on dropout at the club level, as these represent both transfers and people who quit playing (KNVB) organized soccer altogether. Transfers are an interesting case, as these people apparently like the game of soccer, and yet choose to become a member at a different soccer club. The reason for transferring could be related to the social composition of the club. Still, dissatisfaction with the club due to its (changing) social composition, could also lead to KNVB level dropout.

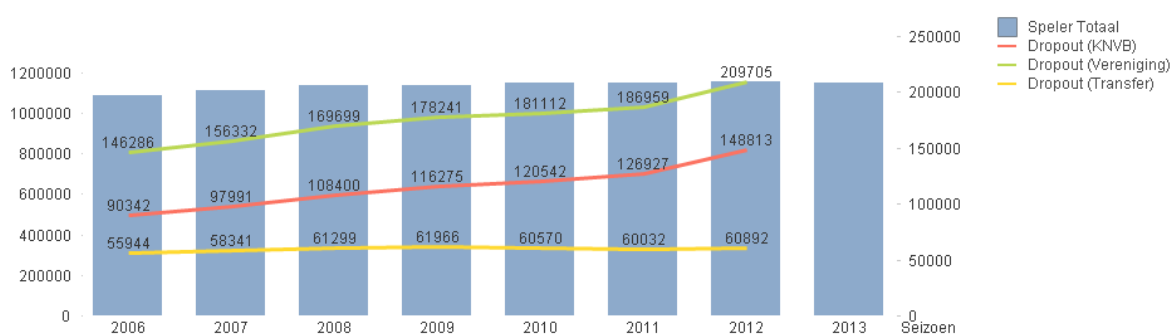


Table 10a. Histogram dropout per season (2006-2013) per dropout type in absolute numbers

Season	Total members	Dropout (club)	Dropout (KNVB)	Dropout (Transfer)
2006-2007	1,085,274	146,286	90,342	55,944
2007-2008	1,113,144	156,332	97,991	58,341
2008-2009	1,134,191	169,699	108,400	61,299
2009-2010	1,135,821	178,241	116,275	61,966
2010-2011	1,145,113	181,112	120,542	60,570
2011-2012	1,149,543	186,959	126,927	60,032
2012-2013	1,155,453	209,705	148,813	60,892

Table 10b. Dropout per season 2006-2013 in absolute numbers

44 | Social Composition & Sport Dropout

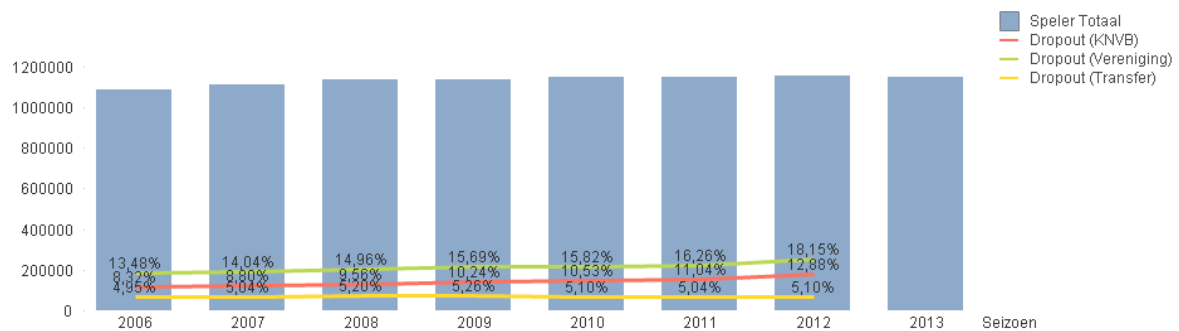


Table 11a. Histogram dropout per season (2006-2013) per dropout type in relative numbers

Season	Total members	Dropout (club)	Dropout (KNVB)	Dropout (Transfer)
2006-2007	1,085,274	13.48%	8.32%	4.95%
2007-2008	1,113,144	14.04%	8.80%	5.04%
2008-2009	1,134,191	14.96%	9.56%	5.20%
2009-2010	1,135,821	15.69%	10.24%	5.26%
2010-2011	1,145,113	15.82%	10.53%	5.10%
2011-2012	1,149,543	16.26%	11.04%	5.04%
2012-2013	1,155,453	18.15%	12.88%	5.10%

Table 11. Dropout per season 2006-2013 in percentages

6. Analyses & Results

In this section the statistical analyses and their results are reported. First, correlations are discussed, both bivariate (at the level of every independent variable separately) and partial (combining certain independent variables). Second, multiple regression is discussed, creating a model for understanding the effects of social composition on dropout. The statistical analyses were done using SPSS 20.0.0.1.

Correlations test the relationship between two variables. In this research, one-tailed tests are executed, as directional expectations were formulated. Also, cases are excluded pairwise, only excluding those cases that do not have a score on the variable necessary for that calculation in particular. In addition, the means and standard deviation were calculated, in order to have an impression of the descriptive data. The assumptions for correlations are partially met: measurements are at least at interval level, either percentages or absolute figures, but a normality distribution cannot be confirmed. Therefore, Spearman's correlation coefficient (r_s) is reported, which is a non-parametric statistic, and can be used when the data violated parametric assumptions, such as non-normally distributed data (Field, 2009). This makes up the dropout & independent variables I section. The independent variables II section is dedicated to partial correlations using bootstrapping, controlling for possible overlap between the variances explained in section I. The direct output is reported in the appendix 2.

Multiple regression is used to create a model that best explains the relationship between the independent variables and dropout. The following configurations were used: six six blocks were entered, the first block contained all the independent variables with stepwise backward method, using all predictors (independent variables) and reassessing the model when one is taken away. The other five blocks contained the different aspects of the independent variables separately using forced entry method. Two plots were requested to assess homoscedasticity and heteroscedasticity on a case-to-case basis. Unknown scores were excluded list wise (most unknown scores were part of the education variable), creating an $N = 2181$ for all variables. During the multiple regression, factor analysis is used to assess if there are latent variables in the independent variables, when results suggested this could be the case. A full description can be found in appendix 3, and the direct output is reported in appendix 4.

6.1 Dropout & independent variables I

A bivariate correlation is a correlation between two variables, however, this correlation says nothing about causality (Field, 2009). Still, looking at correlations tells us something about what to expect from further analyses.

6.1.1 Dropout & ethnicity

Dropout has a mean of .1899 and a standard deviation (SD) of .1751, and the $N = 3738$. Ethnicity is divided in Western and non-Western. Western has a mean of .8899, a SD of .1142, and the $N = 3739$. Non-Western has a mean of .1101, a SD of .1142, and the $N = 3739$.

There is a significant relationship between dropout and being Western, $r_s = -.420$, or being non-Western, $r_s = .420$, all p_s (one-tailed) $< .01$. This significance value tells us that the probability of getting this correlation coefficient, if the null hypothesis were true (there is no relationship between these variables), is very low. However, the r_s indicates that there is only a moderate link between dropout and ethnicity at best. This finding is also evident in the variance explained by ethnicity: R_s^2 is .1764, which means that only 17.64% of the variance is explained. So, although ethnicity is significantly correlated with dropout, it can only account for 17.64% of the variation in dropout.

	Dropout	Western	Non-Western
Dropout	1	-.420	.420
Western	-.420	1	-1
Non-Western	.420	-1	1

Table 1. Correlations (bivariate) dropout & ethnicity, reporting Spearman's rho, all $p_s < .01$

As being Western is negatively correlated with dropout, and being non-Western is positively correlated with dropout, dropout goes down when more Westerners are member, and dropout goes up when more non-Westerners are member. Verweel, Janssens, & Roques (2005) showed that bridging and bonding social capital could play a role in the sport participation of ethnically diverse groups. However, the results of the correlation analysis are in line with what was expected from the hunkering down theory.

6.1.2 Dropout & income

Dropout has a mean of .1899 and a standard deviation (SD) of .1751, and the $N = 3738$.

Income is divided in average, lowest, and highest. Average income has a mean of 2434.48, a SD of 274.46, and the $N = 3739$. Lowest income has a mean of 1199.63, a SD of 333.48, and the $N = 3739$. Highest income has a mean of 5917.33, a SD of 2105.76, and the $N = 3739$.

There is an insignificant relationship between dropout and average income, $r_s = .002$, or highest income, $r_s = -.019$. However there is a significant relationship between dropout and lowest income, $r_s = .092$, p (one-tailed) $< .01$. This significance value tells us that the probability of getting this correlation coefficient, if the null hypothesis were true (there is no relationship between these variables), is very low. However, the r_s indicates that there is only a very weak link between dropout and lowest income at best. This finding is also evident in the variance explained by lowest income: R_s^2 is .0085, which means that only .85% of the variance is explained. So, although lowest income is significantly correlated with dropout, it can only account for .85% of the variation in dropout.

	Dropout	Income (avg)	Income (min)	Income (max)
Dropout	1	.002	.092*	-.019
Income (avg)	.002	1	.235*	.433*
Income (min)	.092*	.235*	1	-.337*
Income (max)	-.019	.433*	-.337*	1

Table 2. Correlations (bivariate) dropout & income, reporting Spearman's rho, * $p < .01$

Minimum is significantly positively correlated with dropout, which is in line with what is expected from theory. Average and maximum income did not have any significant correlation with dropout. Tiessen-Raaphorst, Verbeek, De Haan, & Breedveld (2010) indicated that lack of money is a primary reason for dropout. This is confirmed by the results of the correlation analysis above. However, from theory it could also be derived that having a higher income leads to higher sport participation (Jehoel-Gijsbers, 2004; Tiessen-Raaphorst et al., 2010). This still might be true, but is not due to significantly lower dropout rates in average or maximum income.

6.1.3 Dropout & education

Dropout has a mean of .1899 and a standard deviation (SD) of .1751, and the $N = 3738$. Education is divided in low, medium, and high. Low education has a mean of .4813, a SD of .073, and the $N = 2463$. Medium education has a mean of .3495, a SD of .037, and the $N = 2388$. High education has a mean of .1726, a SD of .0768, and the $N = 2202$.

There is a significant relationship between dropout and education, low education $r_s = -.100$, medium education $r_s = -.181$, and high education $r_s = .182$, all p_s (one-tailed) $< .01$. This significance value tells us that the probability of getting this correlation coefficient, if the null hypothesis were true (there is no relationship between these variables), is very low. However, the r_s indicates that there is only a weak link between dropout and education at best. This finding is also evident in the variance explained by low education: R_s^2 is .01, which means that only 1% of the variance is explained. So, although low education is significantly correlated with dropout, it can only account for 1% of the variation in dropout. The variance explained by medium education: R_s^2 is .033, which means that only 3.3% of the variance is explained. So, although medium education is significantly correlated with dropout, it can only account for 3.3% of the variation in dropout. The variance explained by high education: R_s^2 is .033, which means that only 3.3% of the variance is explained. So, although high education is significantly correlated with dropout, it can only account for 3.3% of the variation in dropout.

	Dropout	Low education	Medium education	High education
Dropout	1	-.100	-.181	.182
Low education	-.100	1	-.227	-.850
Medium education	-.181	-.227	1	-.249
High education	.182	-.850	-.249	1

Table 3. Correlations (bivariate) dropout & education, reporting Spearman's rho, all $p_s < .01$

Education is significantly correlated with dropout. However, contra intuitive results have emerged from the correlation analysis. From theory it was expected that low education would be positively correlated with dropout, and high education would be negatively correlated to dropout (Tiessen-Raaphorst et al., 2010). However, the results show correlations the other way around. In addition, medium education has a stronger negative correlation with dropout than low education, which is also not in line with theory.

6.1.4 Dropout & gender

Dropout has a mean of .1899 and a standard deviation (SD) of .1751, and the $N = 3738$.

Gender is divided in male and female. Male has a mean of .9064, a SD of .1095, and the $N = 3738$. Female has a mean of .0934, a SD of .1095, and the $N = 3738$.

There is a significant relationship between dropout and being male, $r = .291$, or being female, $r = -.292$, all p s (one-tailed) $< .01$. This significance value tells us that the probability of getting this correlation coefficient, if the null hypothesis were true (there is no relationship between these variables), is very low. However, the r s indicates that there is only a weak link between dropout and gender at best. This finding is also evident in the variance explained by gender, R_s^2 is .085 for males, which means that only 8.5% of the variance is explained. So, although being male is significantly correlated with dropout, it can only account for 8.5% of the variation in dropout. The R_s^2 is .085 for females, which means that only 8.5% of the variance is explained. So, although being female is significantly correlated with dropout, it can only account for 8.5% of the variation in dropout.

	Dropout	Male	Female
Dropout	1	.291	-.292
Male	.291	1	-1.000
Female	-.292	-1.000	1

Table 4. Correlations (bivariate) dropout & gender, reporting Spearman's rho, all p s $< .01$

Being male is significantly positively correlated with dropout, and being female is significantly negative correlated with dropout. These results may indicate a break with strong gender stereotyping in sports (Boiche et al., 2013), as more females in a soccer club correlates with less dropout. This is also reflected in the strong increase in female members, while there is a decrease in male members described in the chapter 5 of this thesis. However, to understand the effects of gender on dropout better, an assessment of self-perception and satisfaction (see Boiche et al., 2013; Scanlan et al., 2009) in a selection of the soccer clubs would be in order. This is out of the scope of the current study.

6.1.5 Dropout & age

Dropout has a mean of .1899 and a standard deviation (SD) of .1751, and the $N = 3738$. Age is divided in average, lowest, and highest. Average age has a mean of 28.8, a SD of 6.48, and the $N = 3739$. Lowest age has a mean of 8.46, a SD of 8.83, and the $N = 3738$. Highest age has a mean of 78.09, a SD of 14.36, and the $N = 3738$.

There is a significant relationship between dropout and age, average age $r_s = .167$, lowest age $r_s = .244$, and highest age $r_s = -.259$, all p_s (one-tailed) $<.01$. This significance value tells us that the probability of getting this correlation coefficient, if the null hypothesis were true (there is no relationship between these variables), is very low. However, the r_s indicates that there is only a weak link between dropout and age at best. This finding is also evident in the variance explained by average age: R_s^2 is .0279, which means that only 2.79% of the variance is explained. So, although average age is significantly correlated with dropout, it can only account for 2.79% of the variation in dropout. The variance explained by lowest age: R_s^2 is .0595, which means that only 5.95% of the variance is explained. So, although lowest age is significantly correlated with dropout, it can only account for 5.95% of the variation in dropout. The variance explained by highest age: R_s^2 is .0671, which means that only 6.71% of the variance is explained. So, although highest age is significantly correlated with dropout, it can only account for 6.71% of the variation in dropout.

	Dropout	Age (avg)	Age (min)	Age (max)
Dropout	1	.167	.244	-.259
Age (avg)	.167	1	.471	-.284
Age (min)	.244	.471	1	-.616
Age (max)	-.259	-.284	-.616	1

Table 5. Correlations (bivariate) dropout & age, reporting Spearman's rho, all $p_s <.01$

Average age is positively correlated with dropout. A high maximum age is negatively correlated with dropout, while at the same time a high minimum age is positively correlated to dropout. From theory we would expect that a low minimum age is positively correlated with dropout, as more young people quit playing sports (Tiessen-Raaphorst et al., 2010). A high maximum age is expected to positively correlate with dropout as well, as different age cohorts in adult life have specific reasons to dropout (Tiessen-Raaphorst et al., 2010).

6.1.6 Subconclusion

Dropout and ethnicity had significant but weak correlations, indicating that being Western negatively correlates with dropout, and being non-Western positively correlates with dropout, but this correlation only explains 17.64% of the variance in the data. Dropout and income have negligible positive correlations, except for low income, which was significant but also only explains .85% of the variance in the data. Dropout and education had significant but very weak correlations, low and medium education levels were negatively correlated with dropout, whereas high education level was positively correlated to dropout. Education level explained .9%, 4%, and 3.5% of the variance in the data, respectively. Dropout and gender have significant but weak correlations (male .291, female -.292), explaining 8.5% of the variance. Dropout and age have significant but weak correlations; average age explains 2.79%, minimum age 5.95%, and maximum age 6.71% of the variance in the data. Ethnicity thus explains most of the variance in the data, followed gender and maximum age. These findings are not entirely in line with what we would expect from the data looking at theory, and the expectations formulated. By looking at the Spearman's rho, the parametric assumption of normality was overcome.

6.2 Dropout & independent variables II

A partial correlation is a correlation between two variables in which the effects of other variables are held constant, and partial correlation is used to find out the size of the unique portion of variance (Field, 2009). In this section, first-order partial correlations are calculated for dropout and ethnicity, controlling for education; dropout and income, controlling for education; and dropout and gender, controlling for age. As such a true measure of ethnicity, income, and gender has been obtained, addressing the third variable problem. Partial correlations in SPSS are expressed in Pearson's r , based on the assumption of normality; therefore I performed partial correlations using the bootstrapping method. Bootstrapping is a computationally intensive statistical technique that allows the researcher to make inferences from data without making strong distributional assumptions (Haukoos & Lewis, 2005). In this research bootstrapping is used to overcome assumptions of normality. During the computation of partial correlations, SPSS is directed to take 1000 samples from the data and compute bias corrected and accelerated (BCa) confidence intervals at 95% (see Field, 2012).

6.2.1 Dropout & ethnicity – controlled for education

The outcomes of the partial correlation dropout and ethnicity, controlled for by education are displayed in table 15. First, notice that the partial correlation between dropout and being Western is $-.490$ (BCa 95% CI $-.536, -.440$), which is less than the effect of education is not controlled for ($r = -.533$). Because the BCa 95% CI does not cross zero, we can be confident that the effect in the population is unlikely to be zero and so implies that there is a significant difference between means in the population. Although this correlation is still statistically significant (its p value is still below $.001$), the relationship is diminished. In terms of variance, the value for R^2 for the partial correlation is $.2401$, which means that being Western can now account for 24.01% of the variation in dropout and so the inclusion of education has diminished the amount of variation in dropout shared by being Western, compared to when not controlled for education ($R^2 = .2841$).

Second, notice that the partial correlation between dropout and being non-Western is $.490$ (BCa 95% CI $.433, .541$), which is less than the effect of education is not controlled for ($r = .533$). Because the BCa 95% CI does not cross zero, we can be confident that the effect in the population is unlikely to be zero and so implies that there is a significant difference between means in the population. Although this correlation is still statistically significant (its p value is still below $.001$), the relationship is diminished. In terms of variance, the value for R^2 for the partial correlation is $.2401$, which means that being non-Western can now account for 24.01% of the variation in dropout and so the inclusion of education has diminished the amount of variation in dropout shared by being non-Western, compared to when not controlled for education ($R^2 = .2841$).

Control Variables		Dropout	Western	Non-Western	Low	Medium	High
-none ^a	Dropout	1.000	-.533	.533	-.084	-.216	.187
	Western	-.533	1.000	-1.000	.092	.360	-.264
	Non-Western	.533	-1.000	1.000	-.092	-.360	.264
Low & Medium & High education	Dropout	1.000	-.490	.490			
	Western	-.490	1.000	-1.000			
	Non-Western	.490	-1.000	1.000			

a. Cells contain zero-order (Pearson) correlations.

Table 6. Partial correlation dropout & ethnicity, controlled for education, reporting Pearson's r , $p < .001$

6.2.3 Dropout & income – controlled for education

The outcomes of the partial correlation dropout and income, controlled for by education are displayed in table 16. Notice that the partial correlation between dropout and average income is $-.132$ (BCa 95% CI $-.185, -.074$), which is considerably more than the effect of education is not controlled for ($r = -.052$). In fact the correlation has more than doubled. Because the BCa 95% CI does not cross zero, we can be confident that the effect in the population is unlikely to be zero and so implies that there is a significant difference between means in the population. Although this correlation is still statistically significant (its p value is still below $.001$), the negative relationship is increased. In terms of variance, the value for R^2 for the partial correlation is $.0174$, which means that average income can only account for 1.74% of the variation in dropout and so the inclusion of education has increased the amount of variation in dropout shared by average income, compared to when not controlled for education ($R^2 = .0027$).

Control Variables		Dropout	Income (Avg)	Low	Medium	High
-none ^a	Dropout	1.000	-.052	-.084	-.216	.187
	Income (Avg)	-.052	1.000	-.350	-.102	.384
Low & Medium & High education	Dropout	1.000	-.132			
	Inkomen (Avg)	-.132	1.000			

a. Cells contain zero-order (Pearson) correlations.

Table 7. Partial correlation dropout & income, controlled for education, reporting Pearson's r , $p < .001$

6.2.4 Dropout & gender – controlled for age

The outcomes of the partial correlation dropout and gender, controlled for by age are displayed in table 14. First, notice that the partial correlation between dropout and being male is $.064$ (BCa 95% CI $-.007, .129$), which is considerably less than the effect of age is not controlled for ($r = .090$). In fact the correlation is nearly two-thirds of what it was before. The Bca 95% IC implies that the difference between the means in the population could be negative, positive or even zero. In other words, it is possible that the true difference between means is zero. Therefore, this bootstrap confidence interval confirms that there is a correlation between dropout and being male. Although this correlation is still statistically significant (its p value is still below $.001$), the relationship is diminished. In terms of variance, the value for

R² for the partial correlation is .0041, which means that being male can now only account for .41% of the variation in dropout and so the inclusion of age has severely diminished the amount of variation in dropout shared by being male.

Second, notice that the partial correlation between dropout and being female is -.064 (BCa 95% CI -.122, -.003), which is considerably less than the effect of age is not controlled for ($r = -.091$). In fact the correlation is nearly two-thirds of what it was before. Because the BCa 95% CI does not cross zero, we can be confident that the effect in the population is unlikely to be zero and so implies that there is a significant difference between means in the population. Although this correlation is still statistically significant (its p value is still below .001), the relationship is diminished. In terms of variance, the value for R² for the partial correlation is .0041, which means that being female can now only account for .41% of the variation in dropout and so the inclusion of age has severely diminished the amount of variation in dropout shared by being female.

Control Variables		Dropout	Male	Female	Age (Avg)
-none ^a	Dropout	1.000	.090	-.091	.282
	Male	.090	1.000	-1.000	.105
	Female	-.091	-1.000	1.000	-.105
Age (Avg)	Dropout	1.000	.064	-.064	
	Male	.064	1.000	-1.000	
	Female	-.064	-1.000	1.000	

a. Cells contain zero-order (Pearson) correlations.

Table 8. Partial correlation dropout & gender, controlled for age, reporting Pearson's r, $p < .001$

6.2.5 Subconclusion

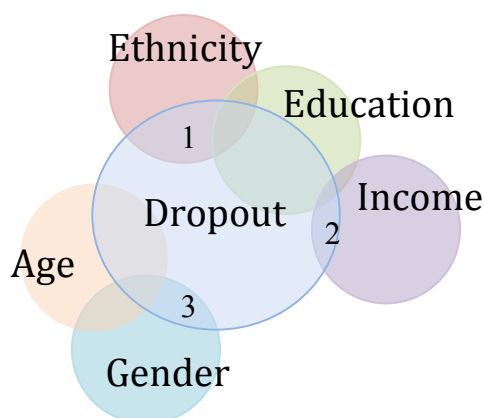


Figure 1. Venn diagram of partial correlation between dropout and independent variables

- 1: Ethnicity explains 24.01% of the variance in dropout when controlled for education.
- 2: Income explains 1.74% of the variance in dropout when controlled for education.
- 3: Gender explains .41% of the variance in dropout when controlled for age.

In this section bootstrapped (1000) partial correlations were discussed between dropout and gender, controlling for age; dropout and ethnicity, controlling for education; and dropout and income, controlling for education. Controlling for education caused the variance in dropout explained by ethnicity alone to decrease from 28.41% (Western) and 28.41% (non-Western) to 24.01% due to changes in the Pearson's r : $-.533$ to $-.490$ (Western) and $.533$ to $.490$ (non-Western). Controlling for education caused the variance in dropout explained by income alone to increase from .27% to 1.74%, due to changes in Pearson's r : $-.052$ to $-.132$.

Controlling for age caused the variance in dropout explained by gender alone to decrease from .81% (male) and .83% (female) to .41% due to changes in the Pearson's r : $.090$ to $.064$ (male) and $-.091$ to $-.064$ (female).

6.3 Conclusion of correlations

In this section I will discuss what correlations implicated overall. Using bivariate correlations and reporting Spearman's rho, the following aspects were interesting with regards to social composition and dropout: ethnicity explains 17.64% of the variance in dropout ($r_s = .420$), lowest income ($r_s = .092$) explains more variance in dropout than highest or average income, low education explains less variance in dropout than medium education ($r_s = -.181$, R_s^2 is 3.3%) or high education ($r_s = .182$, R_s^2 is 3.3%), gender explains 8.5% of the variance in dropout (Male $r_s = .291$, female $r_s = -.292$), and lowest age ($r_s = .244$) and highest age ($r_s = -.259$) explain more variance in dropout than average age, respectively 5.95% and 6.71%.

To understand the correlation between the independent variables and their effect on dropout, (bootstrapped) partial correlations were conducted. This showed a correlation between age, gender, and dropout; between education, ethnicity, and dropout; and education, income, and dropout¹. The first bootstrapped partial correlation caused a decrease in the variance explained by ethnicity from 28.41% to 24.01%. The second bootstrapped partial correlation caused a decrease in the variance explained by income from .27% to 1.74%. This increase could be caused by the way the partial correlation is executed, and point at a moderator, which cannot be researched in this study. The third bootstrapped partial correlation caused an increase in variance explained by gender from .81% to .41%. These examples show that there

¹ These bootstrapped partial correlations report Pearson's r

is shared variance between the independent variables, which should be included in further assessment of expectations. As correlations are not causal relations, in this sense I cannot draw conclusions on the direction of any particular correlations. The expectations cannot be confirmed or rejected on basis of bivariate or partial correlations alone, therefore I will conduct multiple regression in the next section.

6.4 Dropout & social composition

In this section a social composition model is created to predict dropout. First, a summary of the model is given, then the model parameters are given, subsequently the contribution of each independent variable is discussed, next the effects of standard deviation changes are examined and the extreme cases assessed. This section closes with a discussion of the implications of the model.

6.4.1 Summary of the model

The model summary table tells us what the dependent variable (outcome) was and what the predictors were in each model (for a full display of the model summary see appendix 3b). The R^2 is a measure of how much variability in the outcome is accounted for by the predictors. The adjusted R^2 gives us some idea of how well the model generalizes and ideally we would like its value to be very close to R^2 . In this case the difference for the models is small (in fact the difference between values is $.311 - .307 = .004$, about 0.4% maximum). Checking for the Stein's formula for adjusted R^2 gives .302, which is very similar to the observed value for R^2 (.311) indicating that the cross validity of these models is very good. The significance of R^2 can be tested using the F-ratios. Model 1 causes R^2 to change from 0 to .311, and this change in the amount of variance explained gives rise to an F-ratio of 88.948. The addition of new predictors (models 2 to 6) does not cause the F-ratio to change significantly, indicating that the predictors used in model 2 to 6 do not make a large difference. Finally, I requested the Durbin-Watson statistic, this statistic informs us about whether the assumption of independent errors is tenable. For this data the value is 1.964, indicating that the assumption of independent errors has been met.

	R	R ²	Adjusted R ²	Change Statistics					Durbin-Watson
				R ² Change	F Change	df1	df2	Sig. F Change	
1	.558 ^a	.311	.307	.311	88.948	11	2169	.000	
2	.557 ^b	.311	.308	.000	.349	1	2169	.555	
3	.557 ^c	.311	.308	.000	.527	1	2170	.468	
4 ¹	.557 ^d	.310	.308	.000	.883	1	2171	.347	
5	.557 ^e	.311	.307	.000	.475	2	2170	.622	
6	.558 ^f	.311	.307	.000	.780	1	2169	.377	1.964

¹ Predictors: (Constant), Age (Max), Income (Avg), Age (Avg), Non-Western, Income (Min), Income (Max), Age (Min), High Education

Table 9. Model Summary

If the improvement due to fitting the regression model is much greater than the inaccuracy within the model, then the value of F will be greater than 1 and SPSS calculates the exact probability of obtaining the F value by chance. We can interpret these results as meaning that the initial model significantly improved our ability to predict the outcome variable, but that model 4 was even better (because the F-ratio is more significant) (for a full display of the ANOVA table see appendix 3c).

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.054	11	.278	88.948	.000 ^b
	Residual	6.771	2169	.003		
	Total	9.825	2180			
4	Regression	3.049	8	.381	122.153	.000 ^e
	Residual	6.776	2172	.003		
	Total	9.825	2180			

Table 10. ANOVA

6.4.2 Model parameters

So far we have looked at several summary statistics telling us whether or not the models have improved our ability to predict the outcome variable. This part is concerned with the parameters of the model. Model 4 was the best model to predict the outcome variable, and is therefore used in the further analysis (for the results on all models see appendix 3d).

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
4	(Constant)	.190	.016		11.884	.000	.159	.222					
	Non-Western	.280	.012	.470	23.998	.000	.257	.303	.533	.458	.428	.828	1.208
	Income (Avg)	-2.767E-5	.000	-.109	-4.611	.000	.000	.000	-.052	-.098	-.082	.567	1.762
	Income (Min)	-1.204E-5	.000	-.052	-2.396	.017	.000	.000	-.123	-.051	-.043	.671	1.490
	Income (Max)	4.456E-6	.000	.135	5.556	.000	.000	.000	.209	.118	.099	.540	1.851
	High education	.048	.018	.055	2.696	.007	.013	.084	.187	.058	.048	.751	1.331
	Age (Avg)	-.001	.000	-.063	-2.533	.011	-.001	.000	-.092	-.054	-.045	.512	1.954
	Age (Min)	.000	.000	.050	1.670	.095	.000	.001	-.017	.036	.030	.355	2.814
	Age (Max)	.000	.000	-.066	-2.748	.006	-.001	.000	-.066	-.059	-.049	.547	1.828
a. Dependent Variable: Dropout													

Table 11. Coefficients

6.4.3 Independent variables' contribution to the model

The first part of table 11 gives us estimates for the b-values and these values indicate the individual contribution of each predictor in the model, i.e. the values tell us about the relationship between dropout and each predictor. If the value is positive we can tell there is a positive relationship between the predictor and the outcome, whereas a negative coefficient represents a negative relationship. The b-values also tell us to what degree each predictor affects the outcome if the effects of other predictors are held constant. Note that the independent variables reported are based club characteristics and not individual characteristics.

Each of these beta values has an associated standard error indicating to what extent these values would vary across different samples, and these standard errors are used to determine whether or not the b-value differs significantly from zero. If the t-test associated with a b-value is significant then the predictor is making a significant contribution to the model.

For this model being non-Western ($t(2180) = 23.998, p < .001$), average income ($t(2180) = -4.611, p < .001$), maximum income ($t(2180) = 5.556, p < .001$), and maximum age ($t(2180) = -2.748, p < .01$), high education ($t(2180) = 2.696, p < .01$) and minimum income ($t(2180) = -2.396, p < .05$), average age ($t(2180) = -2.533, p < .05$) are significant at the respective significance levels, only ruling out minimum age.

From the magnitude of the t-statistics we can see that the order of impact is as follows (high to low impact):

1. Non-Westerners
2. Maximum income
3. Average income (negative)
4. Maximum age (negative)
5. High education
6. Average age (negative)
7. Minimum income (negative).

6.4.4. Standard deviation change in independent variables and dropout

The standardized betas (table 11) tell us the number of standard deviation that the outcome will change as a result of one standard deviation change in the predictor. The standardized β s make it possible to compare the independent variables. The interpretations are only true if the effects of the other predictors are held constant.

Ethnicity (non-Western) (standardized $\beta = .470$): This value indicates that as being non-Western increases by one standard deviation (.1127), dropout increases by 0.470 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of 0.0315 (0.470×0.0671) in dropout. Therefore for every .1127 increase in non-Westerners, an extra 0.0315 dropout of sports occurs. More non-Western members has a positive effect on dropout, this is in line with what was expected from theory. Theory suggested that a high level of non-Westerners could lead to more dropout, as non-Westerners participate differently in sports (Tiessen-Raaphorst et al., 2010) and the hunkering down effect (see Putnam, 2007) can occur.

Average income (standardized $\beta = -.109$): This value indicates that as average income increases by one standard deviation (€264.64), dropout increases by -0.109 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of -0.0073 ($-.109 \times 0.0671$) in dropout. Therefore for every €264.64 increase in average income, an extra -0.0073 dropout of sports occurs. Average income has a negative effect on dropout, this is not in line with what was expected from theory. Theory suggests that higher (average) income in a club leads to more participation, including a decline in the “don’t have money”-factor that is widely brought up as a reason for dropout (Boiche & Sarrazin, 2009; Tiessen-Raaphorst et al., 2010)

Minimum income (standardized $\beta = -.052$): This value indicates that as minimum income increases by one standard deviation (€290,50), dropout increases by -0.052 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of -0.0035 (-0.052×0.0671) in dropout. Therefore for every €290,50 increase in minimum income, an extra -0.0035 dropout of sports occurs. Minimum income has a small negative effect on dropout, this is in line with what was expected from theory. It was expected from

theory that lower minimum income would lead to higher dropout, as people have less to spend, which is a major reason to dropout (see Jehoel-Gijsbers, 2004).

Maximum income (standardized $\beta = .135$): This value indicates that as maximum income increases by one standard deviation (€2029.50), dropout increases by 0.135 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of 0.0091 (0.135×0.0671) in dropout. Therefore for every €2029.50 increase in maximum income, an extra 0.0091 dropout of sports occurs. Maximum income has a positive effect on dropout, this is not in line with what was expected from theory. It was expected that higher maximum income in a club leads to lower dropout, as participation rates among higher income levels are higher and they have more money to spend on sports (Tiessen-Raaphorst et al., 2010). However, higher maximum income could also indicate that the division between average income levels and the highest income level is too big, creating a hunkering down effect (see Putnam, 2007).

High education (standardized $\beta = .055$): This value indicates that as high education increases by one standard deviation (.0769), dropout increases by 0.055 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of 0.0037 (0.055×0.0671) in dropout. Therefore for every .0769 increase in high education, an extra 0.0037 dropout of sports occurs. High education has a small positive effect on dropout, this is not in line with what was expected from theory. It was expected that high levels of high education income in a club leads to lower dropout, as participation rates among higher education levels are higher and they have more money to spend on sports (Tiessen-Raaphorst et al., 2010). However, higher high education levels could also indicate that the division between low and medium education levels and high education level is too big, creating a hunkering down effect (see Putnam, 2007).

Average age (standardized $\beta = -.063$): This value indicates that as average age increases by one standard deviation (5.71), dropout increases by -0.063 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of -0.0042 (-0.063×0.0671) in dropout. Therefore for every 5.71 years increase in average age, an extra -0.0042 dropout of sports occurs. Average age has a small negative effect on dropout, this is in line with what was expected from theory. Theory suggested that sport participation in general decreases steadily with age (Hendriksen & Hoogwerf, 2013; Tiessen-Raaphorst et al., 2010; Van Bottenburg et al., 2005).

Minimum age (standardized $\beta = .050$): This value indicates that as minimum age increases by one standard deviation (6.96), dropout increases by .050 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of 0.0034 (0.050×0.0671) in dropout. Therefore for every 6.96 years increase in minimum age, an extra 0.0034 dropout of sports occurs. Minimum age has a small positive effect on dropout, this is in line with what was expected from theory. The higher the minimum age in a club, the less likely that younger people (especially aged 12-18 years) influence the dropout rate. 12-18 year-olds make up the age cohort that is most likely to dropout (Fraser-Thomas et al., 2008; Tiessen-Raaphorst et al., 2010).

Maximum age (standardized $\beta = -.066$): This value indicates that as maximum age increases by one standard deviation (12.97), dropout increases by -0.066 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of -0.0044 (-0.066×0.0671) in dropout. Therefore for every 12.97 years increase in maximum age, an extra -0.0044 dropout of sports occurs. Maximum age has a small negative effect on dropout, this is not in line with what was expected from theory. If maximum age increases, the expectation is that dropout increases as well, as older people have other obligations (work and children) or come to face physical constraints in doing sports (Casper, Gray, & Babkes Stellino, 2007; Tiessen-Raaphorst et al., 2010)

6.4.5 Extreme cases

In a sample we would expect 95% of cases to have standardized residuals within +/- 2. The sample used here is 2181, and thus it is expected that about 109 cases (5%) have standardized residuals outside of the limits. The output (see appendix 4f) shows 127 cases (5.82%) that are outside of the limits, therefore the sample is within 1% of what we would expect, which is good. To assess these influential cases we look at the standardized DFBeta values greater than 1, which includes only one case (case 24). The 127 cases should be reassessed in order to make final conclusions about the data, however, due to the limits of this research this will only be done for case 24 (see discussion section).

6.5 Conclusions of social composition model

Ethnicity (Western, non-Western), income (minimum, average, and low), education level (low, medium, and high), gender (male, female), and age (minimum, average, and maximum) were used in a stepwise backward regression to predict dropout. The prediction model (model 4) was statistically significant $F(2180) = 122.153$. $p < .01$, and can account for approximately 31% of the variance of dropout ($R^2 = .310$, adjusted $R^2 = .308$). An overview of the standardized β values for each independent variable is displayed in table 12.

		B	SE Beta	β
4	(Constant)	0.190	0.016	.
	Non-Western	0.280	0.012	.470*
	Income (Avg)	-0.00767	0.000	-.109*
	Income (Min)	-0.00204	0.000	-.052***
	Income (Max)	0.000456	0.000	.135*
	High education	0.048	0.018	.055**
	Age (Avg)	-0.001	0.000	-.063***
	Age (Min)	0.000	0.000	.050
	Age (Max)	0.000	0.000	-.066**

Table 12. Note: $R^2 = .307$ for step 1, $\Delta R^2 = .001$ for step 4 ($p < .001$). * $p < .001$, ** $p < .01$, *** $p < .05$.

The b-values show us the following relationships between independent variables and dropout:

- More non-Western members in a club leads to higher dropout.
- Higher average income of members in a club leads to lower dropout.
- Higher minimum income of members in a club leads to lower dropout.
- Higher maximum income of members in a club leads to higher dropout.
- More people with a high education in a club leads to higher dropout.
- Higher average age of members in a club leads to lower dropout.
- Higher maximum age of members in a club leads to lower dropout.

All these independent variables were significant at $p < .05$ or more stringent. Minimum age and maximum age had no effect on dropout, however maximum age was significant.

The excluded variables list (gender, western, medium and low education) indicated that gender should be researched as well, as it might have an influence on dropout. This aspect also shows in the partial regression plots (see appendix 3g). Due to the limitations of the current research gender cannot be explored further here.

Checking for multicollinearity indicated that the VIF value is at an acceptable level of 1.73, however the tolerance levels are all above 0.2, which could indicate multicollinearity. The collinearity diagnostics for model 4 showed that minimum age might constitute for a problem, and thus should be taken out of the model. This is consistent with the mixed signals of the b and standardized β values for minimum age. Therefore, there can be no final conclusions on this independent variable. Evaluating influential cases pointed out that 127 cases fall outside the ± 2 standardized residuals criterion, which is fine acceptable. The most influential case is case 24, which will be discussed in the conclusion.

The partial regression plots showed that education has very interesting abnormalities, which will be explored in the next section. Gender (male) shows a strong deviance from what is expected if the data is linear and homoscedastic, which indicates that the variability in dropout is not equally distributed across the values of being male. This finding makes sense as high percentages for being male are more common in soccer clubs than low percentages (as less women play soccer). All partial regression plots indicate that the model needs more verification.

To assess the possibility of latent variables I conducted factor analysis, which did not indicate that there were latent variables that influence the fit of the model (see appendix 5). In addition other possible models (assessed in a new multiple regression model) did not change the outcomes (see appendix 6).

7. Conclusion

This research was conducted to assess the relationship between the social composition of soccer clubs and their related dropout rates. The social composition was defined as ethnicity, income, education level, gender, and age. The research uses membership and club data from the KNVB to establish age and gender and different data sources of Statistics Netherlands to estimate the ethnicity, income, and education levels of the members of a soccer club. The current research studied the social composition of soccer clubs and their respective dropout rates and thus tries to fill a gap in the understanding of sport dropout. The basis for studying social composition aspects of a soccer club lies in the theoretical work of Putnam, who discusses bridging and bonding effects in society and identifies a phenomenon which he calls ‘hunkering down.’ In short, hunkering down is behavior in a social context (e.g. a soccer club) that is the outcome of uncertain social expectations driven by diversity, such as diversity in ethnicity, income levels, education levels, gender, or age. Therefore, this research tries to answer the following research question and sub questions:

To what extent does the social composition of a sport club play a role in the decision of members to end their membership?

- 1. What effect has ethnicity on dropout?*
- 2. What effect has income on dropout?*
- 3. What effect has education level on dropout?*
- 4. What effect has gender on dropout?*
- 5. What effect has age on dropout?*

To assess the expectations as outlined in the theoretical model, the different datasets were combined using QlikView. QlikView provided a number of descriptive statistics, and helped to create a table that could be used in SPSS. This table showed the average dropout per club as a percentage; the level of Western and non-Western players in a club, expressed as a percentage; the average income of members of a soccer club, the lowest income of members in euros, and the highest income of members, all expressed in euros; the level of low educated members of a club, the level of medium educated members, and the level of high educated

members, all expressed as a percentage; the gender division of a club (male and female) as a percentage; and the average age of a club, the minimum age of a club, and the maximum age of a club, all expressed in years. All variables were expressed as an average over the seasons 2006-2013 to overcome influences of a 'bad year,' and the unit of analysis was soccer clubs. To do the actual analysis SPSS 20.0.0.1 was used, and correlations were analyzed, multiple regression was conducted (including PCA and a simple regression).

The descriptive statistics that were derived from QlikView informed certain benchmarks. The ethnicity distribution is 89% Western and 11% non-Western among members of soccer clubs. The average income of members of soccer clubs is €2486.61. The education distribution among members of soccer clubs is 47% low, 35% medium, and 18% high. The gender distribution among members of soccer clubs is 90% male, 10% female. The age distribution among members of soccer clubs follows this pattern: soccer is played mostly by youth, starting at the age of 5 and peaking at the age of 14, then rapidly declining until the age of 18, and then slowly declining with a small peak at the age of 45. However, these descriptive statistics say little about the dropout of members of soccer clubs. Therefore, QlikView also created a dropout index, which is .1491 (14.91%) per club on average through seasons 2006-2013. However, the dropout index also shows a steady rise in dropout from season 2006-2007 to 2013-2014.

Correlations between the independent variables and dropout were first analyzed using bivariate correlations (based on Spearman's rho). The partial correlations (expressed in Pearson's r and defined by bootstrapping [1000]) showed that dropout and ethnicity controlled for education caused a decrease in the variance explained by ethnicity alone. Therefore, ethnicity and education are thought to have a correlation, of which the influence on dropout is now established. In addition, the partial correlations analysis showed that dropout and gender controlled for age caused a decrease in the variance explained by income alone. Therefore, gender and age are also thought to have a correlation, of which the influence on dropout is now established. However, the partial correlation analysis showed an interesting deviance from what is expected in the analysis of dropout and income controlled for education: the amount of variance in dropout explained by income increases when controlled for education. This implies that there is third variable that accounts for this change, other than ethnicity of education. The current research cannot research this new (mediating) variable, due to time constraints.

Multiple regression was used to get a sound idea of the expectations. The independent variables were entered into the multiple regression analysis using the backward stepwise method. The model (model 4) that explained dropout the best included non-Western ethnicity, average income, minimum income, maximum income, high education level, average age, minimum age, and maximum age. Gender was not included in the model, but was considered entering into the model. Also, Western ethnicity, low education, and medium education were not entered into the model. For Western ethnicity this makes sense, as it is opposite to non-Western ethnicity. However, for low education and medium education this decision is odd. Consequently, additional factor analysis (PCA) was conducted, creating a latent variable that included low education and high education. Still, a simple regression using this new latent variable indicated that the contribution of low education would be minimal. To understand the influence of education better a new multiple regression was run, only including high education. This new model did not show a significant improvement in explaining dropout.

Below the expectations are evaluated on the bases of correlations and multiple regression, structured by the order in which the sub questions are posed.

7.1 Dropout & ethnicity

E1.1: High percentages of Westerners lead to low dropout rates

E1.2: High percentages of non-Westerners lead to high dropout rates

Ethnicity explains 17.64% of the variance in dropout rates at best, which is relatively low in statistical terms, but in the current research this is the highest value. Being Western is negatively correlated with dropout, which means that if dropout goes up, the percentage of Westerners goes down. For non-Westerners it is just the other way around. Thus, being non-Western and dropout are positively correlated, which means that if dropout goes up, the percentage of non-Westerners goes up as well. These outcomes indicate that these expectations (E1.1 and E1.2) can be confirmed. Multiple regression underpinned these findings by showing that higher levels of non-Westerners leads to higher dropout rates.

7.2 Dropout & income

E2.1: High minimum income leads to low dropout rates

E2.2: High average income leads to low dropout rates

E2.3: High maximum income leads to low dropout rates

Income was divided into average income, lowest income, and highest income, of which only lowest income has a significant correlation with dropout, explaining a mere .85% of the variance. These findings suggest that average and high income are not correlated with dropout. The relevance of this expectation has diminished, but this might change in the light of other variables. Multiple regression showed that higher minimum income and higher average income of members in a club lead to lower dropout, confirming E2.1 and E2.2. However, high maximum income leads to higher dropout rates, contradicting the expectation (E2.3)

7.3 Dropout & education level

E3.1: High percentages of low educated people lead to high dropout rates

E3.2a: High percentages of medium educated people lead to high dropout rates

E3.2b: High percentages of medium educated people lead to low dropout rates

E3.3: High percentages of high educated people lead to low dropout rates

Education was divided into low education, medium education, and high education. All three had a significant correlation with dropout, albeit very low: .76%, 4%, and 3.3%, respectively. This shows that low education level is less correlated with dropout than medium or high education level, still E3.1 should be confirmed. In addition, medium education level explains more variance in dropout than higher education level. These findings suggest that if dropout goes up, the percentage of medium education level goes up as well, confirming expectation E3.2b. The same holds for high education level in a slightly less form, contradicting E3.3. In multiple regression these findings were further assessed; low education and medium education were not included in model 4, indicating that their explanatory value is insignificant. Nonetheless, E3.3 should be rejected, as the model indicated that more people with a high education in a club leads to higher dropout.

7.4 Dropout & gender

E4.1a: High percentages of males lead to low dropout rates

E4.1b: High percentages of males lead to high dropout rates

E4.2a: High percentages of females lead to low dropout rates

E4.2b: High percentages of females lead to high dropout rates

Gender explains 8.5% of the variance in dropout rates at best, which is relatively low in statistical terms, but in the current research this is the second highest value. Being female is negatively correlated with dropout, which means that if dropout goes up, the percentage of females goes down, confirming E4.2a. For males it is just the other way around. Thus, being male and dropout are positively correlated, which means that if dropout goes up, the percentage of males goes up as well, confirming E4.1b. However, on the basis of multiple regression (model 4), these expectations cannot be confirmed; both being male and female are not included in the model.

7.5 Dropout & age

E5.1a: High minimum age leads to low dropout rates

E5.1b: High minimum age leads to high dropout rates

E5.2a: High average age leads to low dropout rates

E5.2b: High average age leads to high dropout rates

E5.3a: High maximum age leads to low dropout rates

E5.3b: High maximum age leads to high dropout rates

The analysis of age is executed on minimum age, average age, and maximum age. Age is significantly and positively correlated with dropout, explaining 5.95%, 2.79%, and 6.71% of the variance in dropout, respectively. These percentages are relatively low in statistical terms, but in the current research these are the third highest values. These findings confirm expectations E5.1b, E5.2b, and E5.3b. In multiple regression (model 4) lowest age and highest age are no longer included in the model. However, the multiple regression model indicates that high average age leads to lower dropout, confirming E5.2a. This finding is contra intuitive from what was found by correlations, but is more accurate as the model contains different variables of social composition.

7.6 Correlations & multiple regression

In tables 13 and 14 these expectations, the values of the correlations, and the direction of the multiple regression are displayed as an adaptation of the theoretical model. Table 13 shows that the correlations indicated 11 correlations between the different variations of the independent variables and dropout. Table 14 shows that the model that resulted from multiple regression is based on 6 of these variations of the independent variables. It is interesting to see that minimum and average income are not included in the correlations, yet are significant in the multiple regression model. What is also striking is that average age is positively correlated with dropout, but is negative in the model, indicating that higher average age reduces dropout. Consistent findings are: higher non-Western ethnicity leads to more dropout, higher minimum income leads to lower dropout, and higher high education levels lead to more dropout.

Independent (high values of)	↔	Outcome
Western ethnicity	-17.64%	Dropout
Non-Western ethnicity	17.64%	
Minimum income	-.85%	
Low education	.76%	
Medium education	4%	
High education	3.3%	
Male	8.5%	
Female	-8.5%	
Minimum age	5.95%	
Average age	2.79%	
High age	6.71%	

Table 13. Significant correlations independent and dropout, expressed in Spearman's Rho

Independent (high values of)	→	Outcome
Non-Western ethnicity	.470	Dropout
Minimum income	-.052	
Average income	-.109	
Maximum income	.135	
High education	.055	
Average age	-.063	
Maximum age	-.066	

Table 14. Significant contribution independent variables to model, expressed in standardized β s

7.7 Dropout & social composition

Returning to the research question, the following can be concluded: Sport participation rates in the Netherlands are still rising (Tiessen-Raaphorst et al., 2010) indicating that the decline of community thesis of Putnam (2001) is not applicable to the Netherlands. However, the phenomena of hunkering down as described by Putnam (2007) might be applicable, as increasingly complex social compositions of sport clubs might lead to higher dropout rates.

This research looked into aspects of dropout that have not been researched before: the social composition effects of the sport club. This blind spot in both literature and practice has been researched using data of the KNVB and Statistics Netherlands, of which the outcomes have been outlined in the results chapter, and combined in this conclusion. The annual amount of dropout for the KNVB (almost 150,000 people per year) alone exceeds the amount of 130,000 people per year indicated by NOC*NSF (Hendriksen & Hoogwerf, 2013). This shows that dropout is an underestimated issue, that is not currently adequately addressed by sport associations and sport clubs.

The outcomes of the current study could inform how to measure dropout in future inquiries in order to get a better picture of the scope of dropout. In addition, this research can be used to inform sport associations and sport clubs on how to understand social composition and their relation to dropout rates in order to create policies that tackle dropout. The current study is also valuable for sport associations and clubs who want to retain low dropout rates. Sport participation, including prolonged membership of sport clubs (i.e. low dropout rates) has important implications for social capital, social cohesion, and health (Hendriksen & Hoogwerf, 2013; Putnam, 2001, 2007; Schnabel et al., 2008; Tiessen-Raaphorst et al., 2010; Verweel et al., 2005). Low dropout rates are not only socially desirable, but also lead to more (paying) members per club, which is beneficial for financial reasons (both for sport clubs and their sport associations).

The answer to the main question is not straightforward. The social composition of a sport club does play a role in the decision of members to end their membership, but there is more to it. Only non-Western ethnicity and high levels of high education actually lead to more dropout. Other social composition aspects lead to lower dropout: higher average income, higher lowest income, higher average age. The social composition model explains 31% of the dropout in sport clubs, the other 69% is still undefined and could be attributable to other reasons. These findings will be discussed in the next section.

8. Discussion

In this section I discuss the concepts and methods used in this research, the results of this research, and formulate interesting directions for future research. In addition, I discuss possible underlying mechanisms of the findings. This section closes with a discussion of homogeneity and heterogeneity as this could be part of the underlying mechanisms, and guide future research into social composition.

8.1 Conceptual discussion

One of the key concepts in this research is social composition. Social composition in itself is not an established concept, therefore social composition was defined in this research in socioeconomic and demographic terms, focusing on the variables ethnicity, income, and education level, gender, and age. The first three variables were estimated using postal codes and data from Statistics Netherlands, rather than data on members of the KNVB. This entails a potential problem as the distribution of these variables in the postal codes are not one-on-one translatable to the individual members of the sport club. Estimating the values on these variables is the closest resemblance to the actual population available as it became clear that the information needed to include ethnicity, income, and education level in the analyses were not available in KNVB datasets. Therefore, new datasets were created, using the alternative of estimating the values of these variables. Another related issue is that the dates related to the point of measurements are not the same for every dataset, which could influence the outcomes. However, these datasets were the best alternative available to assess dropout in relation to social composition.

Another key concept is dropout, which is established using membership data of the KNVB, and is thus directly and adequately relatable to soccer clubs. As soccer clubs are the unit of analysis in this research, dropout was defined as everyone who was not a member of a certain soccer club in the next year (whereas they were in the previous year). This includes all people who either changed soccer clubs (transfers), or who dropped out of soccer (KNVB dropout). For the KNVB as a whole, people who changed soccer clubs might not be

considered dropout, however this research is interested in soccer club related social composition aspects on dropout, and thus includes both types of dropout.

Sport club membership is a much-debated issue in current times as types of membership, and needs and responsibilities of members seem to be changing. One of these changes includes the transition from formal sport participation to informal sport participation, indicating that the sport club might be facing membership loss, while overall sport participation does not decline. The current research only focuses on formal participation in soccer, i.e. playing soccer at a club, and excludes participation in soccer in informal settings (such as Cruijff courts). Informal sportive activities might have a different kind of membership, which could be researched in future studies.

The last conceptual aspect to discuss is a theoretical one. Most literature used is based on sport participation aspects and uses reversed argumentation or logic in order to establish what would cause dropout. Where dropout literature is used, dropout is mostly researched at a personal level, indicating motivations of individuals to dropout (such as I don't have time or money). Dropout based on more aggregate levels, such as the sport club is usually related to adolescents or (former) professional athletes and constitutes sport club aspects such as coaches and atmosphere. These theoretical limitations made it difficult to establish expectations that were entirely theory driven. Therefore, some expectations were formulated in two ways, one expecting lower dropout, and one expecting higher dropout.

8.2 Methodological discussion

The measurement levels are important for data preparation. The choice of measurements as described in the measurement section is based on the idea to create a comparable instrument. The unit of analysis is the sport club, and all sport clubs have received a value for the different independent variables based on the KNVB and Statistics Netherlands data that is comparable (either in percentages or in absolute values). The dependent variable was also measured in percentages to ensure comparability.

For the data analysis two aspects are important: reliability and validity. Reliability of a measure is the extent to which the measure produces the same results when used repeatedly to measure the same thing (Rossi, Lipsey, & Freeman, 2004, p. 218). The reliability of the

measurements derived directly from the KNVB membership data is very high, as this is the actual population data. However, the reliability of the measurements derived from Statistics Netherlands are an estimation of the data in the population (based on the postal codes of the population), and might not be accurate. Still, this was the best alternative (as discussed above). The reliability of the analyses is also assessed by SPSS in the fit of the model to the actual population, which was very high (as is expected from large N datasets). The validity of a measure is the extent to which it measures what it is intended to measure (Rossi et al., 2004, p. 220). Again, the validity of the measurements derived directly from the KNVB membership data is very high, as this is the actual population data. However, for the measurements derived from Statistics Netherlands the validity is harder to establish. The validity is diminished by a couple of aspects: the measurements are estimated and different measurement dates were used. These aspects do not affect the validity of the measurements as such, but do affect the level to which the findings can be generalized. Reliability and validity for this particular data is high, as statistical analyses were done on the whole population rather than a sample.

The methods used in this research are able to contribute to a new way of looking at social composition and dropout. As valuable as this may be, the data was suitable for other analyses as well. One of the possibilities is to conduct longitudinal analysis, looking more in depth into finding portrait in the descriptive analyses of this thesis. Longitudinal analysis would inform us about changes between the years 2006-2013, and thus would give us an idea of what trends are to be found in e.g. social composition over the years.

Another possibility is to conduct survival analysis. This could be done both at the club level as well as on the individual level, to understand changes in social composition and what effects they have on dropout rates. Survival analysis is commonly used in medicine and biology, and typically focuses on time to event data. It consists of techniques for positive random values, such as time to death, time to onset of a disease, but also duration of a strike, or – in this case – time to sport dropout. There are different kinds of survival analysis: clinical trials, prospective cohort studies, retrospective cohort studies, and retrospective correlative studies. The current research would be a form of the latter. Typically, survival data are not fully observed, but rather censored, however, the current study used information about the complete population, rather than a sample and thus censored data would not be needed. Some basic insights into survival analysis are that failure time (T) random variables are always non-negative ($T \geq 0$), T is either discrete or continuous, and a random variable X is called a

censored failure time if $X = \min(T,U)$, where U is a non-negative censoring variable. In the current study this would mean that dropout always follows after the first observation date (season 2006-2007, on April 30, 2007), in this case T is discrete, as observations are made every year, and a random variable is not needed in order to analyze the data. In order to define a failure time random variable the following information is needed: an unambiguous time origin, a time scale, and a definition of the event. In this research the time origin would be season 2006-2007 (April 30, 2007), the time scale is based on real time years, and the event is defined by dropout (i.e. deregistration of membership of a soccer club registered with the KNVB). There are several features typical for survival analysis: individuals do not enter the study at the same time, this is called *staggered entry*, when the study ends some individuals haven't had the event yet, which relates to censoring. Survival analysis is thus one of the possibilities for future analyses.

8.3 Discussion of results

The outcomes of the correlations analysis were mostly in line with what was found by the outcomes of the multiple regression. However, there were more significant correlations between the independent variables and dropout than there were independent variables in model 4. Therefore, both outcomes could help us understand the relation between social composition and dropout.

Ethnicity (being non-Western) explained 17.64% of the variance in dropout, which is in line with what was found in the multiple regression, as being non-Western was still the best predictor of dropout (standardized $\beta = .470$). Of the income variables lowest income had a significant correlation with dropout, and only explained .85% of the variance in dropout. Still, all income variables were included in model 4: average income (standardized $\beta = -.109$) and maximum income (standardized $\beta = .135$) were the following two best predictors of dropout (minimum income had a standardized $\beta = -.052$). Low education had a very weak correlation with dropout (.76%) and medium education had a weak correlation with dropout (4%). Even though high education had a smaller correlation with dropout (3.3%), it was the only education variable included in the social composition model (standardized $\beta = .055$). Other independent variables that were indicated with a relatively high correlation were gender and age. Gender (being male) explained 8.5% of the variance in dropout, but was not included in model 4, although it was considered entering male gender into the model. Lowest age

explained 5.95% of the variance in dropout and highest age explained 6.71% of the dropout, still their standardized β values were low: .050, and -.066 respectively. It would be interesting to research these correlations further, and indicate latent, moderating, or mediating variable(s) that explain(s) the fact that the bivariate correlations and partial correlations differ from expectations, as well as indicate why correlations and multiple regression outcomes differ.

Another interesting aspect found in the data was the influential cases assessment. Case 24 proved most influential in the casewise diagnostics of model 4. The following characteristics belonged to this case: a dropout index of .3804 (38.04%), non-Western members of 10.62%, and average income of €4930.31 (minimum income €2200,- and maximum of €10,000,-), high education of 21%, and an average age of 18.79 years (minimum age 7 years, maximum age 58 years). By looking at these values for the dependent and independent variables, it becomes clear that the dropout percentage is pretty high (compared to an average of 11.1%) and that the average income is also pretty high (compared to an average of €2462.10). This shows in the standardized DFBeta for average income, which is 1.1858 and should be well below 1. Dropout for this case was predicted at 8%, which means that about 30% of the dropout was not predicted by model 4. Therefore the standardized residual value of case 24 is 5.378, which is extreme as we expected this to be +/- 2 or +/- 2.5 tops. To understand better why these values occurred, qualitative analysis of this case (and the 126 other cases that were outside the +/- 2 standardized residuals range) would be in order.

8.4 Recommendations for future research

Some ideas for future research were already mentioned, such as using new methods to analyze the data in a longitudinal way, or conducting survival analyses using the club or individuals as a unit of analysis. This research indicated that there are 127 influential cases, which should be researched in more detail. Analyzing these cases like was done in this discussion with case 24, could give some initial insight into what makes these cases influential. In addition, these cases could be researched using qualitative methods, such as interviews and document analysis (if the club has certain policies). An alternative option would be to research these cases using a survey. An in-depth analysis of these cases would create a better understanding of how social composition variables influence dropout. Another possibility for future research is to look into information in the data that was not researched in

the current study, such as the influence of playing and non-playing members on dropout, the influence of the sport club environment (e.g. neighborhood) on sport club social composition, or look into transfers versus KNVB dropouts. In addition, there are a number of aspects that could influence social composition, which are not included in this research, such as sport club target groups (policy documents and marketing) and image. These aspects could contribute to a more qualitative understanding of social composition, and how social composition could be further operationalized. Dropout could also be understood in terms other than social composition, but still at the level of the sport club (i.e. not related to individual characteristics), such as number of matches played or missed, number of wins, number of yellow or red cards, and so on. This research only used data of one sport (soccer), but it would be very interesting to make similar analyses for other sports, using the same independent variables, to understand differences between different (types of) sports.

8.5 Possible underlying mechanisms

The underlying mechanisms for the outcomes found by the multiple regression in this research are not all straightforward. I discuss high levels of non-Westerners, high income and education, and average age.

When it comes to ethnicity the underlying mechanism might be related to the term hunkering down that is introduced by Putnam (2007): more non-Western members in a club leads to more dropout. This could be due to the fact that sport clubs used to consist out of more Westerners, and the inflow of non-Western members makes the social context of the sport club increasingly complex. However, these results might also be applicable to clubs that were set up by non-Westerners and have a majority of non-Western members. A possible explanation is that the club does have a Western board, but a lot on non-Western members, and both groups have different ideas on how the sport club should be managed (i.e. other ideas on the role of volunteering). In the literature, non-Westerners participate less in organized sports (Tiessen-Raaphorst et al., 2010), and when they do, they might value organized sports differently from Western members leading to higher dropout rates. One issue with defining ethnicity in terms of Westerners and non-Westerners is that it is a generic and empirically non-existing division. Westerners might include people of Dutch descent, but also English or German people, while non-Westerners might include people from Morocco,

Turkey, Surinam, and so on. A major improvement in the findings regarding ethnicity would be to use (or establish) data bases that include more (ethnic) nuances, like described above.

Higher minimum and higher average income leads to lower dropout rates, which is in line with what was expected from theory (Boonstra & Hermens, 2011; Hendriksen & Hoogwerf, 2013; Tiessen-Raaphorst et al., 2010). However, higher maximum income leads to higher dropout rates, and this is contra intuitive. The basic idea is that if people have more to spend, they participate more in sports, and they do this in a more sustainable way. It seems from the data that this basic idea holds up until a certain break point. This break point is related to high incomes. Perhaps people who have more to spend are more critical towards the sport club they are member of. Another possibility is that high incomes are related to frequent travel and moving, causing sport club dropout. Also, hunkering down could be at play here, as a division between people with low and average incomes, and people who earn a lot of money could influence the social composition of the sport club in such a way that the social context is becoming more complex. These could also be explanations of the finding that more people with a high education leads to higher dropout (see e.g. Hendriksen & Hoogwerf, 2013).

The last finding of the multiple regression showed that higher average age of members in a club leads to lower dropout. This is an aspect not specifically described in the literature. One explanation for this finding would be that youth members (12-18 years) are much more likely to dropout than other age cohorts. To assess this idea, the average age of dropouts should be researched per club and then related to the youth cohort. However, this finding also indicates that sport clubs that include more age cohorts (i.e. not only focus on youth players) face fewer dropouts, which leads to a more stable membership base. The stability of the membership base is desirable as it ensures financial means and creates the possibility to achieve a recognizable social context.

8.6 Homogeneity and heterogeneity

The initial idea of this research was to establish homogeneity and heterogeneity of all variables defined in social composition. However, homogeneity and heterogeneity are hard to establish on the basis of the data available. Ideally, these would be formulated using theory driven proof of what homogeneity and heterogeneity are. However, for the variables used to define social composition, this theoretical argument did not exist. Therefore, instead of assessing homogeneity and heterogeneity in effect, the averages and extremes of the variables

were used to create a sense homogeneity and heterogeneity. For the variables ethnicity and gender, these extremes were easily established, as ethnicity was divided in Western and non-Western and gender is either male or female. These characteristics of ethnicity and gender helped to relate the values for these variables per case to the average. The same could be done for age and for income, albeit on three aspects (average, lowest, and highest). However, education proved to be a more difficult case, as this variable was expressed in three different variables that combined added up to 100%. Still, this division could help us understand the influences of education as the average of the three separate variables could be used.

One possibility to research homogeneity and heterogeneity is coined by Coffe & Geys (2007) based on an empirical notion between bonding and bridging social capital (see Putnam, 2000). The crucial aspect in the distinction between bridging and bonding social capital is that they point to different types of socializing (Coffe & Geys, 2007). The extent to which an association is bridging or bonding can be seen as a function of the socioeconomic heterogeneity of its membership. Coffe & Geys (2007) have defined a way to assess a diversity index. They describe that the first step is to calculate the percentage of a given score on a certain function in the population. For example, including all women that are registered at the KNVB. The second step is to calculate the difference between this percentage and the percentage at a given sport club. For example, 20% of all members of the KNVB are have a high education, but only 13% in soccer club X have a high education, concluding to a the difference of -7%. The third step is to calculate the diversity score, in this case the score is equal to the percentage score, but for – for example – categorical income (level) this could be added up and then be divided by the number of categories. The fourth step is to recalculate this score into a diversity index (so-called normalized diversity score) ranging between 0 and 1 per indicator. The last step is to add all the normalized diversity scores to see the overall score. Coffe & Geys (2007) state that in the absence of theoretical arguments the weighing of the different socioeconomic factors should be the same for all factors. In addition, there is some liberty on the side of the researcher to determine the cut off point for bridging and bonding social capital.

For a two-category variable, such as ethnicity (Western or non-Western) or gender (male or female) I use a different, but related way of assessing heterogeneity (bridging) or homogeneity (bonding). This will be outlined in the last chapter of this thesis as the encore.

9. Encore

One of the assumptions is that voluntary organizations, such as sport clubs tend to move towards homogeneity, as recognizable social contexts are important to achieve enduring memberships. To create a better understanding of homogeneity or heterogeneity of independent variables that could be of influence on dropout, this section analyzes ethnicity and gender as indexes using correlations and multiple regression in the same way as in chapter 6 of this thesis. Ethnicity has been researched in the previous sections expressed as percentages of Western and non-Western members of soccer clubs. Gender has been researched in the previous sections expressed as percentages of male and female members of soccer club. In this section the indexes are calculated by taking the highest percentage and subtracting the lowest percentage. For example, a soccer club has 74% Western members and 26% non-Western members, the ethnicity index will be .48 ($74 - 26 = 48$ divided by 100 to create an ethnicity index) or a club has 57% males and 43% females, the gender index will be .13 ($57 - 43 = 13$ divided by 100 to create a gender index). Complete homogeneity (one category is 100%) is thus expressed by the index 1, and complete heterogeneity (both categories are 50%) is expressed by the index 0.

The following expectations were assessed, based on the theoretical framework (chapter 3):

E1: Homogeneity of ethnicity leads to lower dropout rates

E2.1: Homogeneity of gender leads to lower dropout rates

E2.2: Heterogeneity of gender leads to lower dropout rates

9.1 Dropout & indexes I

This section handles bivariate correlations between ethnicity index and dropout, and gender index and dropout. There is a significant relationship between dropout and ethnicity index, $r_s = -.420$, p (one-tailed) $< .01$. This significance value tells us that the probability of getting this correlation coefficient, if the null hypothesis were true (there is no relationship between these

variables), is very low. However, the r_s indicates that there is only a moderate link between dropout and ethnicity index at best. This finding is also evident in the variance explained by ethnicity index: R_s^2 is .1764, which means that only 17.64% of the variance is explained. So, although ethnicity index is significantly correlated with dropout, it can only account for 17.64% of the variation in dropout.

There is a significant relationship between dropout and gender index, $r = .292$, p (one-tailed) $< .01$. This significance value tells us that the probability of getting this correlation coefficient, if the null hypothesis were true (there is no relationship between these variables), is very low. However, the r_s indicates that there is only a weak link between dropout and gender index at best. This finding is also evident in the variance explained by gender index, R_s^2 is .085, which means that only 8.5% of the variance is explained. So, although gender index is significantly correlated with dropout, it can only account for 8.5% of the variation in dropout.

			Dropout	Ethnicity index	Gender index
Spearman's rho	Dropout	Correlation Coefficient	1.000	-.420**	.292**
		Sig. (1-tailed)	.	.000	.000
		N	3737	3737	3737
	Ethnicity index	Correlation Coefficient	-.420**	1.000	-.187**
		Sig. (1-tailed)	.000	.	.000
		N	3737	3738	3737
	Gender index	Correlation Coefficient	.292**	-.187**	1.000
		Sig. (1-tailed)	.000	.000	.
		N	3737	3737	3737

Table 15. Bivariate correlations dropout and indexes, **p (one-tailed) $< .01$

The outcomes of the bivariate correlations are in line with what was reported in the analyses and results chapter of this thesis, thus indicating that ethnicity index and gender index are to be interpreted in the same way as was reported previously.

9.2 Dropout & indexes II

This section handles partial correlations (bootstrapped [1000]) between ethnicity index and dropout controlled for gender, and gender index and dropout controlled for ethnicity. The outcomes of the partial correlation dropout and ethnicity index, controlled for by gender index are displayed in table 16. First, notice that the partial correlation between dropout and ethnicity index is $-.219$ (BCa 95% CI $-.259, -.182$), which is less than the effect of gender index is not controlled for ($r_s = -.420$). Because the BCa 95% CI does not cross zero, we can be confident that the effect in the population is unlikely to be zero and so implies that there is a significant difference between means in the population. Although this correlation is still statistically significant (its p value is still below $.001$), the relationship is diminished. In terms of variance, the value for R^2 for the partial correlation is $.048$, which means that ethnicity index can now account for 4.8% of the variation in dropout and so the inclusion of gender index has diminished the amount of variation in dropout shared by ethnicity index, compared to when not controlled for gender index ($R^2 = .1764$).

Control Variables				Dropout	Ethnicity index	
Gender index	Dropout	Correlation		1.000	-.219	
		Significance (1-tailed)		.	.000	
		df		0	3734	
		Bootstrap ^a	Bias		.000	.000
			Std. Error		.000	.019
			95% Confidence Interval	Lower	1.000	-.259
				Upper	1.000	-.182
	Ethnicity index	Correlation		-.219	1.000	
		Significance (1-tailed)		.000	.	
		df		3734	0	
		Bootstrap ^a	Bias		.000	.000
			Std. Error		.019	.000
			95% Confidence Interval	Lower	-.259	1.000
				Upper	-.182	1.000
a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples						

Table 16. Dropout and ethnicity index, controlling for gender index, reporting Pearson's r

The outcomes of the partial correlation dropout and gender index, controlled for by ethnicity index are displayed in table 17. First, notice that the partial correlation between dropout and

gender index is .165 (BCa 95% CI .129, .200), which is less than the effect of ethnicity index is not controlled for ($r_s = .292$). Because the BCa 95% CI does not cross zero, we can be confident that the effect in the population is unlikely to be zero and so implies that there is a significant difference between means in the population. Although this correlation is still statistically significant (its p value is still below .001), the relationship is diminished. In terms of variance, the value for R^2 for the partial correlation is .0272, which means that gender index can now account for 2.72% of the variation in dropout and so the inclusion of ethnicity index has diminished the amount of variation in dropout shared by gender index, compared to when not controlled for ethnicity index ($R^2 = .085$).

Control Variables				Dropout	Gender index	
Ethnicity index	Dropout	Correlation		1.000	.165	
		Significance (1-tailed)		.	.000	
		df		0	3734	
		Bootstrap ^a	Bias		.000	.000
			Std. Error		.000	.018
			95% Confidence Interval	Lower	1.000	.129
				Upper	1.000	.200
	Gender index	Correlation		.165	1.000	
		Significance (1-tailed)		.000	.	
		df		3734	0	
		Bootstrap ^a	Bias		.000	.000
			Std. Error		.018	.000
			95% Confidence Interval	Lower	.129	1.000
				Upper	.200	1.000

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Table 17. Dropout and gender index, controlling for ethnicity index, reporting Pearson's r

The variance in dropout explained by ethnicity index is reduced from 17.64% to 4.8% when controlled for gender index, and the variance in dropout explained by gender index is reduced from 8.5% to 2.72% when controlled for ethnicity index. This shows that there is a strong overlap between ethnicity index and gender index in explaining dropout.

9.3 Dropout & social composition indexes

In this section the results of multiple regression of the indexes and dropout are discussed. The following configurations were used: one block was entered containing dropout, ethnicity index, and gender index with forced entry method. Two plots were requested to assess homoscedasticity and heteroscedasticity on a case-to-case basis. Unknown scores were excluded list wise, creating an $N = 3737$ for all variables.

9.3.1 Summary of the model

The model summary table tells us what the dependent variable (outcome) was and what the predictors were in the model. The R^2 is a measure of how much variability in the outcome is accounted for by the the predictors. The adjusted R^2 gives us some idea of how well the model generalizes and ideally we would like its value to be very close to R^2 . In this case the difference for the model is small (0.1%). This shrinkage means that if the models were derived from the population, rather than a sample, it would account approximately for 0.1% less variance in the outcome. The significance of R^2 can be tested using F-ratios. Model 1 causes the R^2 to change from 0 to .090, and this change in the amount of variance explained gives rise to an F-ratio of 183.903. Finally, I requested the Durbin-Watson statistic, this informs us about whether the assumption of independent errors is tenable. The closer the value is to 2, the better. For this data the value is 1.785, which is close enough to 2, and thus indicates that the assumption of independent errors almost certainly has been met.

	R	R^2	Adjusted R^2	SE of the Estimate	Change Statistics					Durbin-Watson
					R^2 Change	F Change	df 1	df2	Sig. F Change	
1	.299 ^a	.090	.089	.1671763	.090	183.903	2	3734	.000	1.785

a. Predictors: (Constant), Genderindex, Ethnicityindex

Table 18. Model summary indexes

If the improvement due to fitting the regression model is much greater than the inaccuracy within the model, the value of F will be greater than 1. SPSS calculates the probability of obtaining the F-value by chance (see table 17). For this model the F-ratio is 183.903 which is very unlikely to have happened by chance ($p < .001$).

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.279	2	5.140	183.903	.000 ^b
	Residual	104.357	3734	.028		
	Total	114.637	3736			

Table 19. ANOVA indexes

9.3.2 Model parameters

So far we have looked at several summary statistics telling us whether or not the model has improved our ability to predict the outcome variable. Table 20 is concerned with the parameters of the model.

Coefficients ^a													
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	.188	.020		9.434	.000	.149	.227					
	Ethnicity index	-.186	.014	-.219	-13.702	.000	-.212	-.159	-.253	-.219	-.214	.956	1.046
	Gender index	.179	.018	.164	10.243	.000	.145	.214	.210	.165	.160	.956	1.046

a. Dependent Variable: Dropout

Table 20. Model parameters indexes

9.3.3 Indexes' contribution of the model

The first part of the table gives us estimates for the b-values and these values indicate the individual contribution of each predictor in the model, i.e. the values tell us about the relationship between dropout and each index. If the value is positive, we can tell there is a positive relationship between the predictor and the outcome, whereas a negative coefficient represents a negative relationship. The b-values also tell us to what degree each predictor affects the outcome if the effects of the other indexes are held constant.

Ethnicity index ($b = -0.186$): This value indicates that as ethnicity index increases with one unit, dropout increases with -0.186 units. In other words, every index-point increase in ethnicity, leads to a 0.186% decrease in dropout.

Gender index ($b = 0.179$): This value indicates that as gender index increases with one unit, dropout increases with 0.179 units. In other words, every index-point increase in gender, leads to a 0.179% increase in dropout.

Each of the beta-values has an associated standard error indicating to what extent these values would vary across different samples, and these standard errors are used to determine whether or not the b-value differs significantly from zero. If the t-test associated with a b-value is significant, then the predictor is making a significant contribution to the model. For this model ethnicity index ($t(3736) = -13.702$, $p < .001$) and gender index ($t(3736) = 10.243$) are significant. From the magnitude of the t-statistics we can see that the order of impact is ethnicity index and then gender index.

9.3.4 Standard deviation change in indexes and dropout

The standardized betas tell us the number of standard deviations that the outcome will change as a result of one standard deviation change in the indexes. This interpretation is only true if the other predictor is held constant.

Ethnicity index (standardized $\beta = -.219$): This value indicates that as ethnicity index increases by one standard deviation (.2062), dropout increases by $-.219$ standard deviations.

The standard deviation for dropout is .1752 and so this constitutes a change of -0.0384 (-0.219 x 0.1752) in dropout. Therefore, for every .2062 increase in ethnicity index, a decrease of 0.0384 dropout occurs. This indicates that higher homogeneity in terms of ethnicity causes less dropout.

Gender index (standardized $\beta = .164$): This value indicates that as gender index increases by one standard deviation (.1597), dropout increases by .164 standard deviations. The standard deviation for dropout is .1752 and so this constitutes a change of 0.0287 (0.164 x 0.1752) in dropout. Therefore, for every .1597 increase in gender index, an increase of 0.0287 dropout occurs. This indicates that higher homogeneity in terms of gender causes more dropout.

The sign (positive or negative) of the b-values tells us something about the direction of the relationship between the indexes and dropout. Therefore, we would expect a bad model to have confidence intervals that cross zero. In this model zero is not crossed, and both indexes have a tight confidence interval, indicating that the estimates in the current model are likely to be representative of the true population values.

9.3.5 Extreme cases

In a sample we would expect 95% of cases to have standardized residuals within +/- 2. The sample used here is 3737, and thus it is expected that about 187 cases (5%) have standardized residuals outside of the limits. The output (see appendix 8) shows 169 cases (4.52%) that are outside of the limits.

9.3.6 Checking assumptions

As a final stage in the analysis, the assumptions of the model are checked. To test the normality of residuals, we must look at the histogram and normal probability plot. In a perfectly normally distributed dataset the histogram exactly follows the bell-shaped normal distribution line and all values will follow the straight line in the normal probability plot.

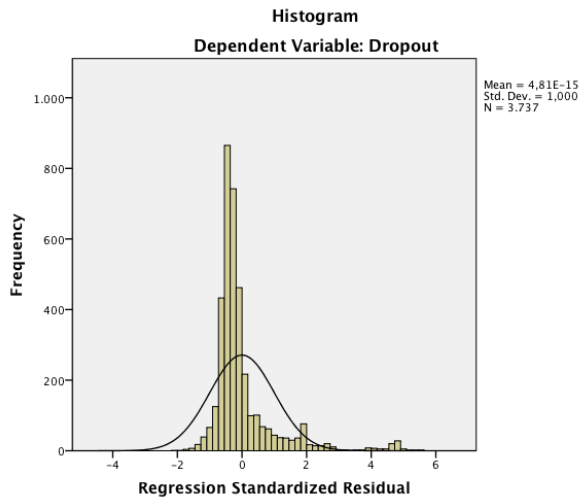


Figure 2. Histogram and normal probability plot of dropout

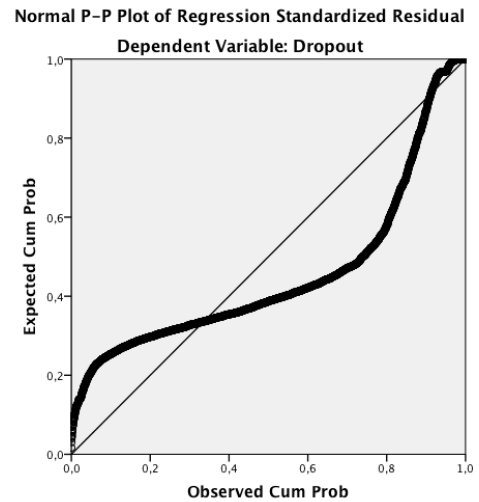


Figure 3. Normal P-Plot of dropout

It is obvious from these figures (2 and 3) that dropout is non-normally distributed. This indicates that the model is not generalizable, however this is not a problem, as the analyses are based on the whole population.

Partial plots were requested, which are scatterplots of the residuals of the outcome variable and each of the predictors (indexes) when both variables are regressed separately on the remaining index. These scatterplots show the relationships (linear / non-linear) and assess homoscedasticity of the indexes and dropout. Homoscedasticity shows if the variability of a variable is equal across the range of values of a second variable that predicts it.

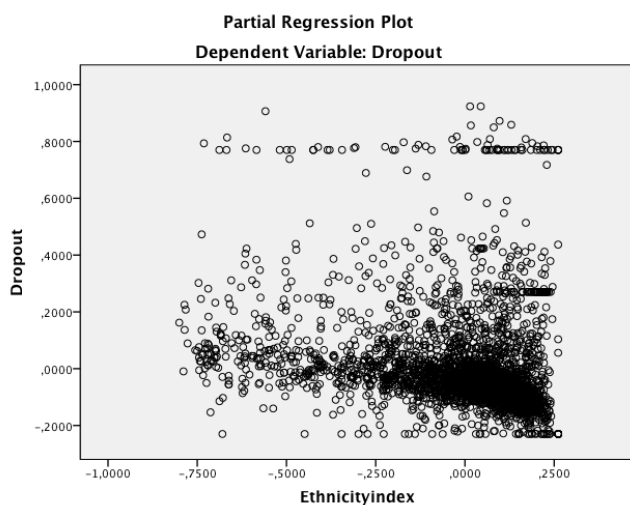


Figure 4. Partial regression plot ethnicity index and dropout

For ethnicity index the partial plot shows a negative relationship to dropout. There are no obvious outliers in this plot. However, the relationship looks like a funnel, indicating heteroscedasticity.

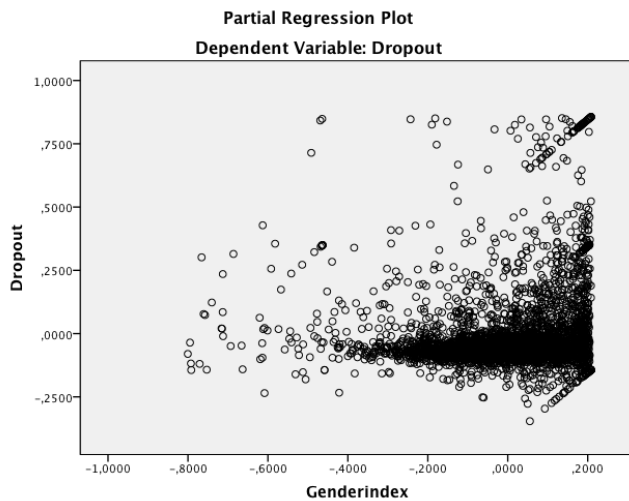


Figure 5. Partial regression plot gender index and dropout

For gender index the partial plot shows a positive relationship to dropout. There are no obvious outliers in this plot. However, the relationship looks like a funnel, indicating heteroscedasticity.

9.4 Conclusions of ethnicity and gender indexes-based model

At the beginning of this chapter, the following expectations were formulated:

E1: Homogeneity of ethnicity leads to lower dropout rates

E2.1: Homogeneity of gender leads to lower dropout rates

E2.2: Heterogeneity of gender leads to lower dropout rates

E1 is confirmed as ethnicity index negatively correlated, indicating homogeneity causes less dropout. E2.2 is confirmed as gender index positively correlated, indicating homogeneity causes more dropout.

This kind of assessment, using indexes, could be expanded to the other social composition variables, creating a measure for homogeneity and heterogeneity of social composition. This is something that could be executed in future research.

Literature

- Archambault, I., Janosz, M., Morizot, J., & Pagani, L. (2009). Adolescent Behavioral, Affective, and Cognitive Engagement in School: Relationship to Dropout. *Journal of School Health, 79*(9), 408–415.
- Barnett, V., & Lewis, T. (1978). *Outliers in statistical data*. New York: Wiley.
- Barreveld, R. (2014). Honkbal moet zich richten op ledenbehoud, niet op groei. *August 13*. Retrieved October 13, 2014, from <http://www.sportenstrategie.nl/2014/sportdeelname/statistieken-en-trends/honkbal-moet-zich-richten-op-ledenbehoud-niet-op-groei/>
- Boiche, J., Plaza, M., Chalabaev, a., Guillet-Descas, E., & Sarrazin, P. (2013). Social Antecedents and Consequences of Gender-Sport Stereotypes During Adolescence. *Psychology of Women Quarterly*.
- Boiche, J., & Sarrazin, P. G. (2009). Proximal and distal factors associated with dropout versus maintained participation in organized sport. *Journal of Sports Science and Medicine, 8*, 9–16.
- Boonstra, N., & Hermens, N. (2011). *De maatschappelijke waarde van sport. De Sportbank / Verwey-Jonker Instituut*. Utrecht.
- Bourdieu, P. (1977). *Outline of a Theory of Practice*. (J. Goody, Ed.) *Cambridge studies in social anthropology* (Vol. 16). Cambridge: Cambridge University Press.
- Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (Vol. 241, pp. 241–258). New York: Greenwood Press.
- Casper, J. M., Gray, D. P., & Babkes Stellino, M. (2007). A Sport Commitment Model Perspective on Adult Tennis Players' Participation Frequency and Purchase Intention. *Sport Management Review, 10*(3), 253–278.
- Coffe, H., & Geys, B. (2007). Toward an Empirical Characterization of Bridging and Bonding Social Capital. *Nonprofit and Voluntary Sector Quarterly, 36*(1), 121–139.
- Collard, D., & Hoekman, R. (2013). *Sportdeelname in Nederland. April*. Utrecht.
- Den Hertog, F., Verweij, A., Mulder, M., Sanderse, C., & Van der Lucht, F. (2014). Sociaaleconomische status: Wat is de huidige situatie? *Volksgezondheid Toekomst Verkenning*. Retrieved October 15, 2014, from <http://www.nationaalkompas.nl/bevolking/sociaaleconomische-status/wat-is-sociaaleconomische-status/>

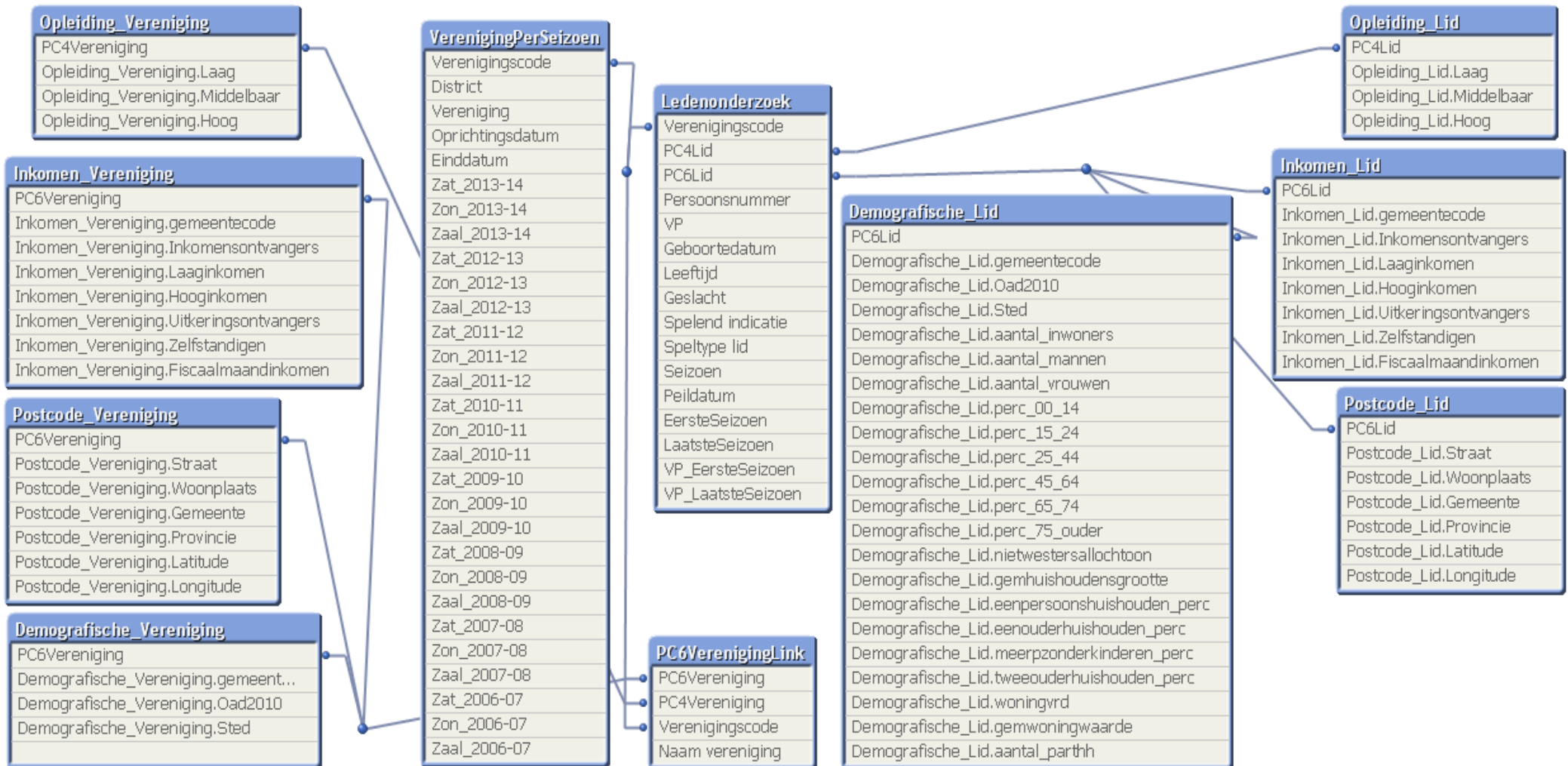
- Elling, A., Knoppers, A., & Knop, P. De. (2001). The Social Integrative Meaning of Sport: A Critical and Comparative Analysis of Policy and Practice in the Netherlands. *Sociology of Sport Journal*, 18, 414–434.
- Explorable. (2009a). Multiple Regression Analysis. Retrieved September 23, 2014, from <https://explorable.com/multiple-regression-analysis>
- Explorable. (2009b). Statistical Correlation. Retrieved September 23, 2014, from <https://explorable.com/statistical-correlation>
- Field, A. (2009). *Discovering Statistics Using SPSS (and sex, drugs, and rock 'n roll)*. London: SAGE.
- Field, A. (2012). Exploring Data : The Beast of Bias Sources of Bias Outliers. In *Discovering Statistics Using SPSS (and sex and drugs and rock "n" roll)* (pp. 1–21).
- Flyvbjerg, B. (2001). *Making Social Science Matter: Why Social Inquiry Fails and How it Can Succeed Again*. Cambridge: Cambridge University Press.
- Fraser-Thomas, Cote, & Deaken. (2008). Examining Adolescent Sport Dropout and Prolonged Engagement from a Developmental Perspective. *Journal of Applied Sport Psychology*, 20(February 2007), 318–333.
- Gucciardi, D. F., & Jackson, B. (2013). Understanding sport continuation: An integration of the theories of planned behaviour and basic psychological needs. *Journal of Science and Medicine in Sport / Sports Medicine Australia*.
- Haukoos, J. S., & Lewis, R. J. (2005). Advanced statistics: bootstrapping confidence intervals for statistics with “difficult” distributions. *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine*, 12(4), 360–5.
- Hendriksen, T., & Hoogwerf, I. (2013). *Sportersmonitor 2012*. Arnhem.
- Hoaglin, D. C., & Welsch, R. E. (1978). The Hat Matrix in Regression and ANOVA. *The American Statistician*, 32(1), 17–22.
- Jehoel-Gijsbers, G. (2004). *Sociale uitsluiting in Nederland*. Den Haag: Sociaal en Cultureel Planbureau.
- Kalmijn, M., & Kraaykamp, G. (2003). Dropout and Downward Mobility in the Educational Career: An Event-History Analysis of Ethnic Schooling Differences in the Netherlands. *Educational Research and Evaluation*, 9(3), 265–287. doi:10.1076/edre.9.3.265.15572
- Kalmthout, J. Van, Jong, M. De, & Lucassen, J. (2009). *Verenigingsmonitor 2008*. W.J.H. Mulier Instituut. Utrecht.
- KNVB. (2014). Who we are. Retrieved October 03, 2014, from <http://english.knvb.nl/whoweare>

- Kuhn, T. S. (1962). *The Structure of Scientific Revolutions* (3rd ed.). Chicago: University of Chicago Press.
- Lepir, D. (2009). Reasons For Withdrawing From Sports In The Formerly Active Athletes. *Physical Culture*, 2(63), 193–203.
- Lim, S. Y., Warner, S., Dixon, M., Berg, B., Kim, C., & Newhouse-bailey, M. (2011). Sport Participation Across National Contexts : A Multilevel Investigation of Individual and Systemic Influences on Adult Sport Participation, (February 2014), 37–41.
- Milne, G. R., & McDonald, M. A. (1999). *Sport Marketing: Managing the Exchange Process* (p. 169). Sudbury: Jones & Bartlett.
- Pharr, S. J., Putnam, R. D., & Dalton, R. J. (2000). A Quarter-Century of Declining Confidence. *Journal of Democracy*, 11(2), 5–25. doi:10.1353/jod.2000.0043
- Plato Stanford. (2014). Models in Science. Retrieved September 23, 2014, from <http://plato.stanford.edu/entries/models-science/>
- Putnam, R. D. (2001). *Bowling Alone*. New York: Simon & Schuster.
- Putnam, R. D. (2007). E Pluribus Unum: Diversity and Community in the Twenty-first Century The 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, 30(2), 137–174.
- Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A Systematic Approach* (Google eBook) (p. 470). New York: SAGE.
- Sawasthi. (2000). Multiple Regression. *University of Texas Arlington*. Retrieved September 29, 2014, from <http://www.uta.edu/faculty/sawasthi/Statistics/stmulreg.html>
- Scanlan, T. K., Russell, D. G., Magyar, T. M., & Scanlan, L. a. (2009). Project on Elite Athlete Commitment (PEAK): III. An examination of the external validity across gender, and the expansion and clarification of the Sport Commitment Model. *Journal of Sport & Exercise Psychology*, 31(6), 685–705.
- Schnabel, P., Bijl, R., & De Hart, J. (2008). *Betrekkelijke betrokkenheid*. Den Haag: SCP.
- Siegle, D. (2014). Correlation Lecture. Retrieved September 29, 2014, from [http://www.gifted.uconn.edu/siegle/research/correlation/correlation notes.htm](http://www.gifted.uconn.edu/siegle/research/correlation/correlation%20notes.htm)
- Siisiäinen, M. (2000). Two Concepts of Social Capital : Bourdieu vs . Putnam.
- Statistics Netherlands. (2014a). Decentrale overheden - postcodegebieden. Statistics Netherlands. Retrieved from <http://www.cbs.nl/nl-NL/menu/informatie/decentrale-overheden/informatie-voor-gemeenten/onderzoek-op-maat-en-voorbeelden/etalage/postcodegebieden/default.htm>

- Statistics Netherlands. (2014b). Toelichting Kerncijfers Postcodegebieden. Statistics Netherlands. Retrieved September 29, 2014, from <http://www.cbs.nl/nr/exeres/faa11085-44e4-4cb1-8481-0315a1b19583>
- Stevens, J. P. (2012). *Applied Multivariate Statistics for the Social Sciences, Fifth Edition*. New York: Routledge.
- Stokvis, R. (1979). *Strijd over sport : organisatorische en ideologische ontwikkelingen I*. Deventer: Van Loghem Slaterus.
- Stuij, M., & Stokvis, R. (2011). Habitusvorming: over de socialisatie van sportgedrag. *Sociologie*, 7(3), 203–222.
- Tiessen-Raaphorst, A., Verbeek, D., De Haan, J., & Breedveld, K. (2010). *Sport : een leven lang*. Den Haag/Den Bosch: SCP / W.J.H.Mulier Instituut.
- Tsai, P. K., & Gracy, R. W. (1976). Isolation and characterization of crystalline methylglyoxal synthetase from *Proteus vulgaris*. *The Journal of Biological Chemistry*, 251(2), 364–7.
- Van Bottenburg, M., Rijnen, B., & Van Sterkenburg, J. (2005). *Sport Participation in the EU: trends and differences*. Amsterdam.
- Van Ingen, E. (2009). *Let's Come Together and Unite. Studies of the Changing Character of Voluntary Association Participation*. Tilburg University.
- Van Sterkenburg, J. (2012). Race/ethnicity, sport and the research/policy relationship. The Dutch context. *Journal of Policy Research in Tourism, Leisure and Events*, 4(1), 112–116.
- Verweel, P., Janssens, J., & Roques, C. (2005). Kleurrijke zuilen: Over de ontwikkeling van sociaal kapitaal in eigen en gemengde sportverenigingen. *Vrijtijdsstudies*, 23(4), 7–22.

Appendices

1. QlikView model



2. Correlations - output

a. Bivariate

I Ethnicity & dropout

Descriptive Statistics			
	Mean	Std. Deviation	N
Dropout	,18993267607509	,175151358491622	3738
Westers	,88988420622451	,114228584184498	3739
Niet Westers	,11011579377549	,114228584184498	3739

Correlations					
			Dropout	Westers	Niet Westers
Spearman's rho	Dropout	Correlation Coefficient	1,000	-,420**	,420**
		Sig. (1-tailed)	.	,000	,000
		N	3738	3738	3738
	Westers	Correlation Coefficient	-,420**	1,000	-1,000**
		Sig. (1-tailed)	,000	.	.
		N	3738	3739	3739
	Niet Westers	Correlation Coefficient	,420**	-1,000**	1,000
		Sig. (1-tailed)	,000	.	.
		N	3738	3739	3739

** . Correlation is significant at the 0.01 level (1-tailed).

II Income & dropout

Descriptive Statistics			
	Mean	Std. Deviation	N
Dropout	,18993267607509	,175151358491622	3738
Inkomen (Avg)	2434,48020789594330	274,456721827228250	3739
Inkomen (Min)	1199,62556833377900	333,480228118036200	3739
Inkomen (Max)	5917,33083712222600	2105,757665443415600	3739

Correlations						
			Dropout	Inkomen (Avg)	Inkomen (Min)	Inkomen (Max)
Spearman's rho	Dropout	Correlation Coefficient	1,000	,002	,092**	-,019
		Sig. (1-tailed)	.	,453	,000	,122
		N	3738	3738	3738	3738
	Inkomen (Avg)	Correlation Coefficient	,002	1,000	,235**	,433**
		Sig. (1-tailed)	,453	.	,000	,000
		N	3738	3739	3739	3739
	Inkomen (Min)	Correlation Coefficient	,092**	,235**	1,000	-,337**
		Sig. (1-tailed)	,000	,000	.	,000
		N	3738	3739	3739	3739
	Inkomen (Max)	Correlation Coefficient	-,019	,433**	-,337**	1,000
		Sig. (1-tailed)	,122	,000	,000	.
		N	3738	3739	3739	3739

** . Correlation is significant at the 0.01 level (1-tailed).

III Education & dropout

Descriptive Statistics			
	Mean	Std. Deviation	N
Dropout	,18993267607509	,175151358491622	3738
Laag	,48130540930441	,072988750826424	2463
Middelbaar	,34946907782539	,038682609793467	2388
Hoog	,17261393811214	,076824479861109	2202

Correlations						
			Dropout	Laag	Middelbaar	Hoog
Spearman's rho	Dropout	Correlation Coefficient	1,000	-,100**	-,181**	,182**
		Sig. (1-tailed)	.	,000	,000	,000
		N	3738	2462	2387	2201
	Laag	Correlation Coefficient	-,100**	1,000	-,227**	-,850**
		Sig. (1-tailed)	,000	.	,000	,000
		N	2462	2463	2382	2200
	Middelbaar	Correlation Coefficient	-,181**	-,227**	1,000	-,249**
		Sig. (1-tailed)	,000	,000	.	,000
		N	2387	2382	2388	2183
	Hoog	Correlation Coefficient	,182**	-,850**	-,249**	1,000
		Sig. (1-tailed)	,000	,000	,000	.
		N	2201	2200	2183	2202

** . Correlation is significant at the 0.01 level (1-tailed).

IV Gender & dropout

Descriptive Statistics			
	Mean	Std. Deviation	N
Dropout	,18993267607509	,175151358491622	3738
Man	,90642401315654	,109460324989432	3738
Vrouw	,09343254693822	,109486592442442	3738

Correlations					
			Dropout	Man	Vrouw
Spearman's rho	Dropout	Correlation Coefficient	1,000	,291**	-,292**
		Sig. (1-tailed)	.	,000	,000
		N	3738	3738	3738
	Man	Correlation Coefficient	,291**	1,000	-1,000**
		Sig. (1-tailed)	,000	.	,000
		N	3738	3738	3738
	Vrouw	Correlation Coefficient	-,292**	-1,000**	1,000
		Sig. (1-tailed)	,000	,000	.
		N	3738	3738	3738
** . Correlation is significant at the 0.01 level (1-tailed).					

V Age & dropout

Descriptive Statistics			
	Mean	Std. Deviation	N
Dropout	,18993267607509	,175151358491622	3738
Leeftijd (Avg)	28,80320684003339	6,480777047562154	3739
Leeftijd (Min)	8,45960406634564	8,828888048495660	3738
Leeftijd (Max)	78,09176029962546	14,357449158022003	3738

Correlations						
			Dropout	Leeftijd (Avg)	Leeftijd (Min)	Leeftijd (Max)
Spearman's rho	Dropout	Correlation Coefficient	1,000	,167**	,244**	-,259**
		Sig. (1-tailed)	.	,000	,000	,000
		N	3738	3738	3738	3738
	Leeftijd (Avg)	Correlation Coefficient	,167**	1,000	,471**	-,284**
		Sig. (1-tailed)	,000	.	,000	,000
		N	3738	3739	3738	3738
	Leeftijd (Min)	Correlation Coefficient	,244**	,471**	1,000	-,616**
		Sig. (1-tailed)	,000	,000	.	,000
		N	3738	3738	3738	3738
	Leeftijd (Max)	Correlation Coefficient	-,259**	-,284**	-,616**	1,000
		Sig. (1-tailed)	,000	,000	,000	.
		N	3738	3738	3738	3738

** . Correlation is significant at the 0.01 level (1-tailed).

b. Partial - bootstrapped

1. Gender & dropout – controlled for age

Descriptive Statistics						
		Statistic	Bootstrap ^a			
			Bias	Std. Error	BCa 95% Confidence Interval	
					Lower	Upper
Dropout	Mean	,18993267607509	,00003440525250	,00294186870019	,18435262099395	,19578022919328
	Std. Deviation	,175151358491622	-,000063995708574	,004768429381021	,165967080085101	,184452611705281
	N	3738	0	0	.	.
Vrouw	Mean	,09343254693822	,00000048367116	,00176627149471	,08994892046166	,09691223871958
	Std. Deviation	,109486592442442	-,000287820360203	,004629624417091	,100736055312182	,118079188477544
	N	3738	0	0	.	.
Man	Mean	,90642401315654	-,00000069424466	,00176585108308	,90294801497594	,90990982401660
	Std. Deviation	,109460324989432	-,000287937335370	,004630446838375	,100713103563016	,118043626311417
	N	3738	0	0	.	.
Leeftijd (Avg)	Mean	28,80368173728204	-,00011222621083	,10722865636464	28,59678940849951	29,01862001668440
	Std. Deviation	6,481579032568556	-,006683562634811	,114508470811916	6,266274338648063	6,686498024816394
	N	3738	0	0	.	.

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Correlations							
Control Variables			Dropout	Vrouw	Man	Leeftijd (Avg)	
-none ^a	Dropout	Correlation	1,000	-,091	,090	,282	
		Significance (1-tailed)	.	,000	,000	,000	
		df	0	3736	3736	3736	
		Bootstrap ^b	Bias	,000	-,001	,001	,000
			Std. Error	,000	,031	,031	,022
BCa 95% Confidence	Lower		.	-,150	,021	,239	

			Interval	Upper	.	-,029	,156	,325		
	Vrouw	Correlation				-,091	1,000	-1,000	-,105	
		Significance (1-tailed)				,000	.	,000	,000	
		df				3736	0	3736	3736	
		Bootstrap ^b	Bias				-,001	,000	,000	-,002
			Std. Error				,031	,000	,000	,025
	BCa 95% Confidence Interval	Lower			-,150	.	-1,000	-,151		
		Upper			-,029	.	-1,000	-,061		
	Man	Correlation				,090	-1,000	1,000	,105	
		Significance (1-tailed)				,000	,000	.	,000	
		df				3736	3736	0	3736	
Bootstrap ^b		Bias				,001	,000	,000	,002	
		Std. Error				,031	,000	,000	,025	
BCa 95% Confidence Interval	Lower			,021	-1,000	.	,052			
	Upper			,156	-1,000	.	,161			
Leeftijd (Avg)	Correlation				,282	-,105	,105	1,000		
	Significance (1-tailed)				,000	,000	,000	.		
	df				3736	3736	3736	0		
	Bootstrap ^b	Bias				,000	-,002	,002	,000	
		Std. Error				,022	,025	,025	,000	
BCa 95% Confidence Interval	Lower			,239	-,151	,052	.			
	Upper			,325	-,061	,161	.			
Leeftijd (Avg)	Dropout	Correlation				1,000	-,064	,064		
		Significance (1-tailed)				.	,000	,000		
		df				0	3735	3735		
		Bootstrap ^b	Bias				,000	-,001	,001	
			Std. Error				,000	,031	,031	
	BCa 95% Confidence Interval	Lower			.	-,122	-,007			
		Upper			.	-,003	,129			
	Vrouw	Correlation				-,064	1,000	-1,000		
		Significance (1-tailed)				,000	.	,000		
		df				3735	0	3735		
Bootstrap ^b		Bias				-,001	,000	,000		
	Std. Error				,031	,000	,000			

			BCa 95% Confidence Interval	Lower	-,122	.	-1,000		
				Upper	-,003	.	-1,000		
	Man	Correlation			,064	-1,000	1,000		
		Significance (1-tailed)			,000	,000	.		
		df			3735	3735	0		
		Bootstrap ^b	Bias			,001	,000	,000	
			Std. Error			,031	,000	,000	
	BCa 95% Confidence Interval		Lower	-,007	-1,000	.			
				Upper	,129	-1,000	.		
a. Cells contain zero-order (Pearson) correlations.									
b. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples									

II. Ethnicity & dropout – controlled for education

Bootstrap Specifications	
Sampling Method	Simple
Number of Samples	1000
Confidence Interval Level	95,0%
Confidence Interval Type	Bias-corrected and accelerated (BCa)

Descriptive Statistics						
		Statistic	Bootstrap ^a			
			Bias	Std. Error	BCa 95% Confidence Interval	
					Lower	Upper
Dropout	Mean	,12988739670150	,00002973844300	,00142343726460	,12715616636842	,13292337542954
	Std. Deviation	,067133322205874	-,000063559133497	,001424097806443	,064292882329355	,069728222717533
	N	2181	0	0	.	.
Westers	Mean	,89213760438983	-,00000669437567	,00234076434222	,88763827611312	,89632033527499

	Std. Deviation	,112728987257569	-,000133841673940	,003351704897641	,106722201444429	,119000156263473
	N	2181	0	0	.	.
Niet Westers	Mean	,10786239561017	,00000669437567	,00234076434222	,10320560156669	,11262069728262
	Std. Deviation	,112728987257569	-,000133841673940	,003351704897641	,106722201444429	,119000156263473
	N	2181	0	0	.	.
Laag	Mean	,47852283529773	,00007139453345	,00155443347968	,47533631652627	,48184737538779
	Std. Deviation	,072450305258876	-,000157356678324	,001692975849107	,068982697314333	,075244090508127
	N	2181	0	0	.	.
Middelbaar	Mean	,34864376539549	-,00000555140595	,00079205650088	,34702782928029	,35027900041984
	Std. Deviation	,037754259732394	-,000045659386320	,000741775737067	,036277856289748	,039118964319983
	N	2181	0	0	.	.
Hoog	Mean	,17302892492954	-,00006612682114	,00166810808296	,16983167596797	,17621408248672
	Std. Deviation	,076889207011452	-,000138227808044	,002010972582105	,072814053468239	,080275416770208
	N	2181	0	0	.	.

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Correlations										
Control Variables				Dropout	Westers	Niet Westers	Laag	Middelbaar	Hoog	
-none ^a	Dropout	Correlation		1,000	-,533	,533	-,084	-,216	,187	
		Significance (1-tailed)		.	,000	,000	,000	,000	,000	
		df		0	2179	2179	2179	2179	2179	
		Bootstrap ^b	Bias		,000	,000	,000	,002	-,001	-,002
			Std. Error		,000	,025	,025	,026	,023	,024
			BCa 95% Confidence Interval	Lower	.	-,577	,481	-,135	-,259	,140
				Upper	.	-,488	,583	-,027	-,174	,228
	Westers	Correlation		-,533	1,000	-1,000	,092	,360	-,264	
		Significance (1-tailed)		,000	.	,000	,000	,000	,000	
		df		2179	0	2179	2179	2179	2179	
		Bootstrap ^b	Bias		,000	,000	,000	,000	-,001	,001
			Std. Error		,025	,000	,000	,033	,022	,026
			BCa 95% Confidence Interval	Lower	-,577	.	-1,000	,027	,316	-,317
				Upper	-,488	.	-1,000	,157	,401	-,207

		Interval									
Niet Westers	Correlation			,533	-1,000	1,000	-,092	-,360	,264		
	Significance (1-tailed)			,000	,000	.	,000	,000	,000		
	df			2179	2179	0	2179	2179	2179		
	Bootstrap ^b	Bias			,000	,000	,000	,000	,001	-,001	
		Std. Error			,025	,000	,000	,033	,022	,026	
		BCa 95% Confidence Interval	Lower		,481	-1,000	.	-,158	-,403	,215	
			Upper		,583	-1,000	.	-,026	-,313	,312	
	Laag	Correlation			-,084	,092	-,092	1,000	-,139	-,872	
		Significance (1-tailed)			,000	,000	,000	.	,000	,000	
		df			2179	2179	2179	0	2179	2179	
		Bootstrap ^b	Bias			,002	,000	,000	,000	,001	,000
			Std. Error			,026	,033	,033	,000	,035	,006
BCa 95% Confidence Interval			Lower		-,135	,027	-,158	.	-,211	-,883	
			Upper		-,027	,157	-,026	.	-,071	-,859	
Middelbaar		Correlation			-,216	,360	-,360	-,139	1,000	-,358	
	Significance (1-tailed)			,000	,000	,000	,000	.	,000		
	df			2179	2179	2179	2179	0	2179		
	Bootstrap ^b	Bias			-,001	-,001	,001	,001	,000	-,001	
		Std. Error			,023	,022	,022	,035	,000	,028	
		BCa 95% Confidence Interval	Lower		-,259	,316	-,403	-,211	.	-,409	
			Upper		-,174	,401	-,313	-,071	.	-,307	
	Hoog	Correlation			,187	-,264	,264	-,872	-,358	1,000	
Significance (1-tailed)			,000	,000	,000	,000	,000	.			
df			2179	2179	2179	2179	2179	0			
Bootstrap ^b		Bias			-,002	,001	-,001	,000	-,001	,000	
		Std. Error			,024	,026	,026	,006	,028	,000	
		BCa 95% Confidence Interval	Lower		,140	-,317	,215	-,883	-,409	.	
			Upper		,228	-,207	,312	-,859	-,307	.	
Laag &		Dropout	Correlation	1,000	-,490	,490					

Middelbaar & Hoog		Significance (1-tailed)		.	,000	,000				
		df		0	2176	2176				
		Bootstrap ^b	Bias		,000	-,001	,001			
			Std. Error		,000	,026	,026			
			BCa 95% Confidence Interval	Lower	.	-,536	,433			
				Upper	.	-,440	,541			
		Westers	Correlation		-,490	1,000	-1,000			
	Significance (1-tailed)		,000	.	,000					
	df		2176	0	2176					
	Bootstrap ^b		Bias		-,001	,000	,000			
			Std. Error		,026	,000	,000			
			BCa 95% Confidence Interval	Lower	-,536	.	-1,000			
				Upper	-,440	.	-1,000			
	Niet Westers	Correlation		,490	-1,000	1,000				
		Significance (1-tailed)		,000	,000	.				
		df		2176	2176	0				
		Bootstrap ^b	Bias		,001	,000	,000			
			Std. Error		,026	,000	,000			
			BCa 95% Confidence Interval	Lower	,433	-1,000	.			
				Upper	,541	-1,000	.			
a. Cells contain zero-order (Pearson) correlations.										
b. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples										

III. Income & dropout – controlled for education

Bootstrap Specifications	
Sampling Method	Simple

Number of Samples	1000
Confidence Interval Level	95,0%
Confidence Interval Type	Bias-corrected and accelerated (BCa)

Descriptive Statistics						
		Statistic	Bootstrap ^a			
			Bias	Std. Error	BCa 95% Confidence Interval	
					Lower	Upper
Dropout	Mean	,12988739670150	-,00000918998035	,00145562921061	,12719936970640	,13271039484192
	Std. Deviation	,067133322205874	- ,000056403578486	,001372805754019	,064345279794532	,069773795236323
	N	2181	0	0	.	.
Inkomen (Avg)	Mean	2440,61284980366870	-,04788484655910	5,76669935471102	2429,58721800402100	2452,45501738160330
	Std. Deviation	264,638489976578600	- ,114431374107198	7,908856494537996	251,520138239520380	279,827289320393560
	N	2181	0	0	.	.
Inkomen (Min)	Mean	1164,87849610270520	,06822558459453	6,35107210744557	1151,80578302823400	1177,17706301276640
	Std. Deviation	290,499169258922100	,023026023809052	6,368642934462138	278,220169411614000	303,256184371261300
	N	2181	0	0	.	.
Inkomen (Max)	Mean	6087,57450710683100	,68358551124220	43,98602616549972	5998,40011491132200	6180,48124257427700
	Std. Deviation	2029,490521001774400	- ,664731252541060	25,625222928795754	1978,509006435163000	2082,885397619889000
	N	2181	0	0	.	.
Laag	Mean	,47852283529773	,00004927661036	,00155917086993	,47532298692580	,48184447945722
	Std. Deviation	,072450305258876	,000026671694936	,001662771828756	,069069298184698	,075733021772773
	N	2181	0	0	.	.
Middelbaar	Mean	,34864376539549	-,00002231075908	,00079625340016	,34711295793410	,35021578890055
	Std. Deviation	,037754259732394	- ,000000837693514	,000744028559278	,036182521498767	,039191994834634
	N	2181	0	0	.	.
Hoog	Mean	,17302892492954	-,00002754445546	,00168558079155	,16998658378789	,17638489118712

	Std. Deviation	,076889207011452	- ,000024450707385	,001950728967170	,073250601277249	,080569074104462
	N	2181	0	0	.	.

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Correlations											
Control Variables				Dropout	Inkomen (Avg)	Inkomen (Min)	Inkomen (Max)	Laag	Middelbaar	Hoog	
-none ^a	Dropout	Correlation		1,000	-,052	-,123	,209	-,084	-,216	,187	
		Significance (1-tailed)		.	,008	,000	,000	,000	,000	,000	
		df		0	2179	2179	2179	2179	2179	2179	
		Bootstrap ^b	Bias		,000	-,001	-,001	,000	,000	-,001	,001
			Std. Error		,000	,027	,023	,021	,026	,021	,025
			BCa 95% Confidence Interval	Lower	.	-,097	-,169	,164	-,135	-,257	,140
		Upper		.	-,003	-,080	,252	-,034	-,176	,236	
	Inkomen (Avg)	Correlation		-,052	1,000	,265	,470	-,350	-,102	,384	
		Significance (1-tailed)		,008	.	,000	,000	,000	,000	,000	
		df		2179	0	2179	2179	2179	2179	2179	
		Bootstrap ^b	Bias		-,001	,000	,000	,000	,001	,002	-,001
			Std. Error		,027	,000	,024	,017	,023	,026	,026
			BCa 95% Confidence Interval	Lower	-,097	.	,219	,435	-,396	-,157	,335
		Upper		-,003	.	,313	,505	-,303	-,044	,430	
	Inkomen (Min)	Correlation		-,123	,265	1,000	-,181	-,058	-,001	,057	
		Significance (1-tailed)		,000	,000	.	,000	,003	,474	,004	
		df		2179	2179	0	2179	2179	2179	2179	
		Bootstrap ^b	Bias		-,001	,000	,000	,000	,000	,000	,000
			Std. Error		,023	,024	,000	,022	,021	,023	,021
			BCa 95% Confidence Interval	Lower	-,169	,219	.	-,222	-,099	-,045	,014
		Upper		-,080	,313	.	-,140	-,017	,043	,100	

Inkomen (Max)	Correlation		,209	,470	-,181	1,000	-,235	-,215	,333	
	Significance (1-tailed)		,000	,000	,000	.	,000	,000	,000	
	df		2179	2179	2179	0	2179	2179	2179	
	Bootstrap ^b	Bias		,000	,000	,000	,000	,001	,000	-,001
		Std. Error		,021	,017	,022	,000	,022	,021	,021
		BCa 95% Confidence Interval	Lower	,164	,435	-,222	.	-,282	-,254	,293
	Upper		,252	,505	-,140	.	-,189	-,173	,371	
	Laag	Correlation		-,084	-,350	-,058	-,235	1,000	-,139	-,872
		Significance (1-tailed)		,000	,000	,003	,000	.	,000	,000
		df		2179	2179	2179	2179	0	2179	2179
Bootstrap ^b		Bias		,000	,001	,000	,001	,000	-,001	,000
		Std. Error		,026	,023	,021	,022	,000	,033	,006
		BCa 95% Confidence Interval	Lower	-,135	-,396	-,099	-,282	.	-,202	-,883
Upper			-,034	-,303	-,017	-,189	.	-,078	-,859	
Middelbaar	Correlation		-,216	-,102	-,001	-,215	-,139	1,000	-,358	
	Significance (1-tailed)		,000	,000	,474	,000	,000	.	,000	
	df		2179	2179	2179	2179	2179	0	2179	
	Bootstrap ^b	Bias		-,001	,002	,000	,000	-,001	,000	,001
		Std. Error		,021	,026	,023	,021	,033	,000	,027
		BCa 95% Confidence Interval	Lower	-,257	-,157	-,045	-,254	-,202	.	-,408
Upper	-,176		-,044	,043	-,173	-,078	.	-,299		
Hoog	Correlation		,187	,384	,057	,333	-,872	-,358	1,000	
	Significance (1-tailed)		,000	,000	,004	,000	,000	,000	.	
	df		2179	2179	2179	2179	2179	2179	0	
	Bootstrap ^b	Bias		,001	-,001	,000	-,001	,000	,001	,000
		Std. Error		,025	,026	,021	,021	,006	,027	,000
		BCa 95% Confidence Interval	Lower	,140	,335	,014	,293	-,883	-,408	.
Upper	,236		,430	,100	,371	-,859	-,299	.		
Laag & Middelbaar &	Dropout	Correlation		1,000	-,132	-,135	,142			
		Significance (1-tailed)		.	,000	,000	,000			

Hoog		df	0	2176	2176	2176				
	Bootstrap ^b	Bias	,000	-,001	-,001	,000				
		Std. Error	,000	,030	,023	,022				
		BCa 95% Confidence Interval	Lower	.	-,185	-,181	,094			
			Upper	.	-,074	-,090	,186			
	Inkomen (Avg)	Correlation	-,132	1,000	,263	,399				
		Significance (1-tailed)	,000	.	,000	,000				
		df	2176	0	2176	2176				
		Bootstrap ^b	Bias	-,001	,000	,000	,001			
	Std. Error		,030	,000	,024	,019				
	BCa 95% Confidence Interval		Lower	-,185	.	,214	,363			
			Upper	-,074	.	,312	,438			
	Inkomen (Min)	Correlation	-,135	,263	1,000	-,213				
		Significance (1-tailed)	,000	,000	.	,000				
		df	2176	2176	0	2176				
		Bootstrap ^b	Bias	-,001	,000	,000	,000			
	Std. Error		,023	,024	,000	,021				
	BCa 95% Confidence Interval		Lower	-,181	,214	.	-,253			
			Upper	-,090	,312	.	-,172			
	Inkomen (Max)	Correlation	,142	,399	-,213	1,000				
Significance (1-tailed)		,000	,000	,000	.					
df		2176	2176	2176	0					
Bootstrap ^b		Bias	,000	,001	,000	,000				
	Std. Error	,022	,019	,021	,000					
	BCa 95% Confidence Interval	Lower	,094	,363	-,253	.				
		Upper	,186	,438	-,172	.				

a. Cells contain zero-order (Pearson) correlations.

b. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

3. Multiple regression – full description

a. Descriptives

The correlations table (see appendix 3a) shows that of all the predictors the percentage of ethnicity correlates best with the outcome ($r = .533$, $p < .001$). These and other outcomes in this table are also assessed in the correlations section above. Also, it can be established that there is no multicollinearity in the data, as there are no substantial correlations ($r > .9$) between predictors, however this will be assessed in more detail below.

b. Summary of the model

The model summary table tells us what the dependent variable (outcome) was and what the predictors were in each model (for a full display of the model summary see appendix 3b). The R^2 is a measure of how much variability in the outcome is accounted for by the predictors. The adjusted R^2 gives us some idea of how well the model generalizes and ideally we would like its value to be very close to R^2 . In this case the difference for the models is small (in fact the difference between values is $.311 - .307 = .004$, about 0.4% maximum). This shrinkage means that if the models were derived from the population, rather than a sample, it would account approximately for 0.4% less variance in the outcome. Checking for the Stein's formula for adjusted R^2 gives .302, which is very similar to the observed value for R^2 (.311) indicating that the cross validity of these models is very good. The significance of R^2 can be tested using the F-ratios. Model 1 causes R^2 to change from 0 to .311, and this change in the amount of variance explained gives rise to an F-ratio of 88.948. The addition of new predictors (models 2 to 6) does not cause the F-ratio to change significantly, indicating that the predictors used in model 2 to 6 do not make a large difference. Finally, I requested the Durbin-Watson statistic, this statistic informs us about whether the assumption of independent errors is tenable. The closer the value is to 2, the better, for this data the value is 1.964, which is very close to 2, and thus indicates that the assumption of independent errors almost certainly has been met.

	R	R ²	Adjusted R ²	Change Statistics					Durbin-Watson
				R ² Change	F Change	df1	df2	Sig. F Change	
1	,558 ^a	,311	,307	,311	88,948	11	2169	,000	
2	,557 ^b	,311	,308	,000	,349	1	2169	,555	
3	,557 ^c	,311	,308	,000	,527	1	2170	,468	
4 ¹	,557 ^d	,310	,308	,000	,883	1	2171	,347	
5	,557 ^e	,311	,307	,000	,475	2	2170	,622	
6	,558 ^f	,311	,307	,000	,780	1	2169	,377	1,964

¹ Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog

Table 21. Model Summary

If the improvement due to fitting the regression model is much greater than the inaccuracy within the model, then the value of F will be greater than 1 and SPSS calculates the exact probability of obtaining the F value by chance. For model 1 the F-ratio is 88.948, which is very unlikely to have happened by chance ($p < .001$). For model 2 the F-ratio is 97.837, which is very unlikely to have happened by chance ($p < .001$). For model 3 the F-ratio is 108.673, which is very unlikely to have happened by chance ($p < .001$). For model 4 the F-ratio is 122.153, which is very unlikely to have happened by chance ($p < .001$). For model 5 the F-ratio is 97.770, which is very unlikely to have happened by chance ($p < .001$). For model 6 the F-ratio is 88.944, which is very unlikely to have happened by chance ($p < .001$). We can interpret these results as meaning that the initial model significantly improved our ability to predict the outcome variable, but that model 4 was even better (because the F-ratio is more significant) (for a full display of the ANOVA table see appendix 3c).

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,054	11	,278	88,948	,000 ^b
	Residual	6,771	2169	,003		
	Total	9,825	2180			
2	Regression	3,053	10	,305	97,837	,000 ^c
	Residual	6,772	2170	,003		
	Total	9,825	2180			
3	Regression	3,052	9	,339	108,673	,000 ^d
	Residual	6,773	2171	,003		
	Total	9,825	2180			
4	Regression	3,049	8	,381	122,153	,000 ^e
	Residual	6,776	2172	,003		
	Total	9,825	2180			
5	Regression	3,052	10	,305	97,770	,000 ^f
	Residual	6,773	2170	,003		
	Total	9,825	2180			
6	Regression	3,054	11	,278	88,944	,000 ^g
	Residual	6,771	2169	,003		
	Total	9,825	2180			

Table 22. ANOVA

c. Model parameters

So far we have looked at several summary statistics telling us whether or not the models have improved our ability to predict the outcome variable. This part is concerned with the parameters of the model. Model 4 was the best model to predict the outcome variable, and is therefore used in the further analysis (for the results on all models see appendix 3d).

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
4	(Constant)	,190	,016		11,884	,000	,159	,222					
	Non-Western	,280	,012	,470	23,998	,000	,257	,303	,533	,458	,428	,828	1,208
	Income (Avg)	-2,767E-5	,000	-,109	-4,611	,000	,000	,000	-,052	-,098	-,082	,567	1,762
	Income (Min)	-1,204E-5	,000	-,052	-2,396	,017	,000	,000	-,123	-,051	-,043	,671	1,490
	Income (Max)	4,456E-6	,000	,135	5,556	,000	,000	,000	,209	,118	,099	,540	1,851
	High education	,048	,018	,055	2,696	,007	,013	,084	,187	,058	,048	,751	1,331
	Age (Avg)	-,001	,000	-,063	-2,533	,011	-,001	,000	-,092	-,054	-,045	,512	1,954
	Age (Min)	,000	,000	,050	1,670	,095	,000	,001	-,017	,036	,030	,355	2,814
	Age (Max)	,000	,000	-,066	-2,748	,006	-,001	,000	-,066	-,059	-,049	,547	1,828

a. Dependent Variable: Dropout

Table 23. Coefficients

The first part of the table gives us estimates for the b-values and these values indicate the individual contribution of each predictor in the model, i.e. the values tell us about the relationship between dropout and each predictor. If the value is positive we can tell there is a positive relationship between the predictor and the outcome, whereas a negative coefficient represents a negative relationship. The b-values also tell us to what degree each predictor affects the outcome if the effects of other predictors are held constant.

Ethnicity (non-Western) ($b = 0.280$): This value indicates that as being non-Western increases with one unit, dropout increases by 0.280 units. Both variables were measured as percentages; therefore every percentage-point increase in non-Westerners leads to a 0.28% increase in dropout. This interpretation is only true if the other predictors (described below) are held constant.

Average income ($b = -0.00002767$): This value indicates that as average income increases with one unit, dropout increases by -0.00002767 units. Average income was measured in euros (absolute numbers); therefore every euro increase in average income leads to a -0.002767% increase in dropout. This interpretation is only true if the other predictors are held constant.

Minimum income ($b = -0.00001204$): This value indicates that as minimum increases with one unit, dropout increases by -0.00001204 units. Minimum income was measured in euros (absolute numbers); therefore every euro increase in average income leads to a -0.001204% increase in dropout. This interpretation is only true if the other predictors are held constant.

Maximum income ($b = 0.000004456$): This value indicates that as maximum income increases with one unit, dropout increases by 0.000004456 units. Maximum income was measured in euros (absolute numbers); therefore every euro increase in maximum income leads to a 0.000004456% increase in dropout. This interpretation is only true if the other predictors are held constant.

High education ($b = 0.048$): This value indicates that as high education increases with one unit, dropout increases by 0.048 units. Both variables were measured as percentages; therefore every percentage-point increase in high education leads to a 0.048% increase in dropout. This interpretation is only true if the other predictors are held constant.

Average age ($b = -.001$): This value indicates that as average age increases with one unit, dropout decreases by -0.001 units. Average age was measured in years (absolute numbers); therefore a year increase in average age leads to a 0.001% decrease in dropout. This interpretation is only true if the other predictors are held constant.

Minimum age ($b = 0.000$): This value indicates that as minimum age increases with one unit, dropout increases by 0.000 units. Minimum age was measured in years (absolute numbers); therefore a year increase in minimum age leads to a 0.000% increase in dropout. This interpretation is only true if the other predictors are held constant.

Maximum age ($b = 0.000$): This value indicates that as maximum age increases with one unit, dropout increases by 0.000 units. Maximum age was measured in years (absolute numbers); therefore a year increase in maximum age leads to a 0.000% increase in dropout. This interpretation is only true if the other predictors are held constant.

Each of these beta values has an associated standard error indicating to what extent these values would vary across different samples, and these standard errors are used to determine whether or not the b-value differs significantly from zero. If the t-test associated with a b-value is significant then the predictor is making a significant contribution to the model.

For this model being non-Western ($t(2180) = 23.998, p < .001$), average income ($t(2180) = -4.611, p < .001$), maximum income ($t(2180) = 5.556, p < .001$), and maximum age ($t(2180) = -2.748, p < .01$), high education ($t(2180) = 2.696, p < .01$) and minimum income ($t(2180) = -2.396, p < .05$), average age ($t(2180) = -2.533, p < .05$) are significant at the respective significance levels, only ruling out minimum age. From the magnitude of the t-statistics we can see that the order of impact is as follows (high to low impact): being non-Western, maximum income, average income (negative), maximum age (negative), high education, average age (negative), and minimum income (negative).

The standardized betas tell us the number of standard deviation that the outcome will change as a result of one standard deviation change in the predictor.

Ethnicity (non-Western) (standardized $\beta = .470$): This value indicates that as being non-Western increases by one standard deviation ($.1127$), dropout increases by 0.470 standard

deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of 0.0315 (0.470×0.0671) in dropout. Therefore for every .1127 increase in non-Westerners, an extra 0.0315 dropout of sports. This interpretation is only true if the effects of the other predictors are held constant.

Average income (standardized $\beta = -.109$): This value indicates that as average income increases by one standard deviation (€264.64), dropout increases by -0.109 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of -0.0073 ($-.109 \times 0.0671$) in dropout. Therefore for every €264.64 increase in average income, an extra -0.0073 dropout of sports. This interpretation is only true if the effects of the other predictors are held constant.

Minimum income (standardized $\beta = -.052$): This value indicates that as minimum income increases by one standard deviation (€290,50), dropout increases by -0.052 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of -0.0035 (-0.052×0.0671) in dropout. Therefore for every €290,50 increase in minimum income, an extra -0.0035 dropout of sports. This interpretation is only true if the effects of the other predictors are held constant.

Maximum income (standardized $\beta = .135$): This value indicates that as maximum income increases by one standard deviation (€2029.50), dropout increases by 0.135 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of 0.0091 (0.135×0.0671) in dropout. Therefore for every €2029.50 increase in maximum income, an extra 0.0091 dropout of sports. This interpretation is only true if the effects of the other predictors are held constant.

High education (standardized $\beta = .055$): This value indicates that as high education increases by one standard deviation (.0769), dropout increases by 0.055 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of 0.0037 (0.055×0.0671) in dropout. Therefore for every .0769 increase in high education, an extra 0.0037 dropout of sports. This interpretation is only true if the effects of the other predictors are held constant.

Average age (standardized $\beta = -.063$): This value indicates that as average age increases by one standard deviation (5.71), dropout increases by -0.063 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of -0.0042 (-0.063 x 0.0671) in dropout. Therefore for every 5.71 years increase in average age, an extra -0.0042 dropout of sports. This interpretation is only true if the effects of the other predictors are held constant.

Minimum age (standardized $\beta = .050$): This value indicates that as minimum age increases by one standard deviation (6.96), dropout increases by .050 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of 0.0034 (0.050 x 0.0671) in dropout. Therefore for every 6.96 years increase in minimum age, an extra 0.0034 dropout of sports. This interpretation is only true if the effects of the other predictors are held constant.

Maximum age (standardized $\beta = -.066$): This value indicates that as maximum age increases by one standard deviation (12.97), dropout increases by -0.066 standard deviations. The standard deviation for dropout is 0.0671 and so this constitutes a change of -0.0044 (-0.066 x 0.0671) in dropout. Therefore for every 12.97 years increase in maximum age, an extra -0.0044 dropout of sports. This interpretation is only true if the effects of the other predictors are held constant.

The sign (positive or negative) of the b values tells us something about the direction of the relationship between the predictor and the outcome. Therefore, we would expect a bad model to have confidence intervals that cross zero. In this model average age and maximum age cross the zero in the confidence interval, all other predictors do not cross the zero confidence interval and are positive. All predictors have a tight confidence interval indicating that the estimates in the current model are likely to be representative of the true population values.

d. Excluded variables

In a stepwise regression the excluded variables table contains a summary of all the variables that SPSS is considering entering into the model (for the complete output see appendix 3e). The summary gives an estimate of each predictors beta value if it was entered into the equation at this point and calculates a t-test for this value. SPSS should enter the predictor with the highest t-statistic and will continue to enter predictors until there are none left with

significance values less than .05. The partial correlation provides an indication as to what contribution an excluded predictor would make if it were entered into the model (Field, 2009). In a stepwise regression SPSS should enter the predictor with the highest t-statistic, in this case gender, which is significant ($p < .05$). All other important variables are already in the models.

Model		Beta In	t	Sig.	Partial Correlation
1	Westers	. ^b	.	.	.
	Man	-4,675 ^b	-2,024	,043	-,043
2	Westers	. ^c	.	.	.
	Man	-4,659 ^c	-2,018	,044	-,043
	Middelbaar	,069 ^c	,590	,555	,013
3	Westers	. ^d	.	.	.
	Man	-4,640 ^d	-2,010	,045	-,043
	Middelbaar	-,012 ^d	-,614	,539	-,013
	Laag	,028 ^d	,726	,468	,016
4	Westers	. ^e	.	.	.
	Man	,017 ^e	,924	,356	,020
	Middelbaar	-,013 ^e	-,666	,505	-,014
	Laag	,030 ^e	,778	,437	,017
	Vrouw	-,017 ^e	-,940	,347	-,020
5	Westers	. ^f	.	.	.
	Man	,016 ^f	,883	,377	,019
	Vrouw	-,016 ^f	-,899	,369	-,019
6	Westers	. ^g	.	.	.
	Vrouw	-4,692 ^g	-2,031	,042	-,044

Table 24. Excluded variables

e. Assessing the assumption of multicollinearity

In table 19 (coefficients) collinearity statistics were reported: VIF and tolerance. There are a couple of guidelines for assessing these outcomes: the largest VIF should not be greater than 10, if the average VIF is substantially larger than 1 the regression might be biased, tolerance below 0.2 indicates a problem. For model 4 the VIF values are well below 10. The average VIF is 1.73, which is close enough to 1, and thus confirms that collinearity is not a problem for this model. However, the tolerance statistics are all above 0.2; therefore we cannot conclude that there is no multicollinearity within the data. Table 21 (collinearity diagnostics)

shows the Eigenvalues and the variance proportions of model 4. To assess collinearity we look at large variance proportions and small Eigenvalues. The variance proportions vary between 0 and 1, and for each predictor should be distributed across different dimensions (or Eigenvalues). For model 4 you can see that almost each predictor has most of its variance loading onto a different dimension (non-Western has 43% variance on dimension 2, and 37% variance on dimension 3, high education has 73% variance on dimension 4, maximum income has 55% variance on dimension 5, minimum income has 71% variance on dimension 6, average age has 84% variance on dimension 7, maximum age has 68% variance on dimension 8, income has 70% of variance on dimension 9). Minimum age is divided (29% variance on dimension 7, and 23% variance on dimension 9). For minimum age these outcomes are not unexpected as there could be a relation between average and minimum age, however, there is no cause for concern as the variance proportion of minimum age is substantially lower than average age and maximum age respectively.

Dimension	Eigen value	Condition Index	Variance Proportions									
			(Constant)	Non-Western	Income Avg	Income Min	Income Max	High	Age Avg	Age Min	Age Max	
1	7,779	1,000	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
2	,535	3,814	,00	,43	,00	,00	,00	,00	,00	,00	,12	,00
3	,425	4,278	,00	,37	,00	,00	,01	,00	,00	,00	,17	,00
4	,132	7,688	,00	,09	,00	,01	,01	,73	,01	,00	,00	,01
5	,066	10,859	,00	,02	,00	,07	,55	,18	,01	,15	,00	,00
6	,036	14,795	,00	,00	,00	,71	,07	,01	,06	,04	,06	,06
7	,015	22,588	,03	,01	,05	,01	,18	,00	,84	,29	,08	,08
8	,009	29,017	,07	,00	,25	,19	,03	,05	,05	,23	,68	,06
9	,004	46,783	,89	,08	,70	,01	,15	,03	,03	,00	,16	,16

Table 25. Collinearity diagnostics model 4

f. Casewise diagnostics

In a sample we would expect 95% of cases to have standardized residuals within +/- 2. The sample used here is 2181, and thus it is expected that about 109 cases (5%) have standardized residuals outside of the limits. The output (see appendix 3f) shows 127 cases (5.82%) that are outside of the limits, therefore the sample is within 1% of what we would expect, which is good. In addition, 99% of the cases should lie within +/- 2.5. From the cases in the SPSS

output it is clear that 60 (2.75%) lie outside these limits, which is reason to further investigate this output. Appendix 3f shows the case summaries for all the 127 cases that that fall outside the +/- 2 standardized residuals. The case summaries table shows that there are no cases with a Cook's distance greater than 1 (the highest number is 0,16208 for case 24), which is thus no cause for concern. The average leverage can be calculated as p/n , for this data that is 0.0073, and so we are looking for values of $2(p/n) = 0.015$, following Hoaglin & Welsh (1978) or for values of $3(p/n) = 0.022$, following Stevens (Stevens, 2012). There are respectively 16 and 5 cases that exceed these cut-off points, however, cases with large leverage values will not have a large influence on the regression coefficients per se. Therefore, the Mahalanobis distances are also assessed, which measure the distance of cases from the means of the predictor variables. From Barnett & Lewis (1978) it can be derived that with large samples values above 25 are cause for concern. In the current data this comes down to 19 cases. To assess these influential cases we look at the standardized DFBeta values greater than 1, which includes only one case (case 24). See the table below for a summary of the most influential cases.

	Case	InkomenAvg	Cook	Mahalanobis	Leverage
1	13	0,03066	0,02531	28,35	0,013
2	24	1,18575	0,16208	128,69754	0,05904
3	151	0,07605	0,01899	42,75268	0,01961
4	526	0,07049	0,0064	31,21575	0,01432
5	637	-0,07492	0,00846	33,67306	0,01545
6	695	-0,09761	0,01944	99,27491	0,04554
7	942	-0,04554	0,00913	32,93064	0,01511
8	1064	0,06482	0,00742	35,57002	0,01632
9	1138	0,01838	0,0123	31,44175	0,01442
10	1379	0,09008	0,01822	41,04527	0,01883
11	1470	0,07966	0,04943	34,18439	0,01568
12	2416	0,05951	0,00586	34,47022	0,01581
13	2581	-0,14593	0,00893	33,23675	0,01525
14	2610	-0,05701	0,02384	34,68216	0,01591
15	2715	0,05536	0,08791	66,07432	0,03031
16	2753	-0,0704	0,03629	32,85143	0,01507
17	3036	-0,15664	0,00915	48,62299	0,0223
18	3040	-0,03049	0,01758	63,2367	0,02901
19	3501	-0,03405	0,02291	102,91482	0,04721

Table 26. Case summaries most influential cases

The 127 cases, and especially those listed above should be reassessed in order to make final conclusions about the data, however, due to the limits of this research this will only be done for case 24 (see discussion section).

g. Checking assumptions

As a final stage in the analysis, the assumptions of the model are checked. To test the normality of the residuals, we must look at the histogram and normal probability plot. In a perfectly normally distributed dataset the histogram exactly follows the bell-shaped normal distribution line and all values will follow the straight line in the normal probability plot.

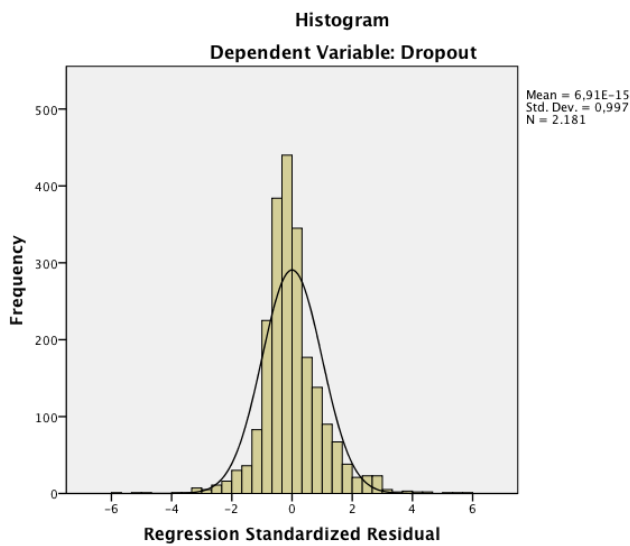


Figure 3. Histogram and normal probability plot of dropout

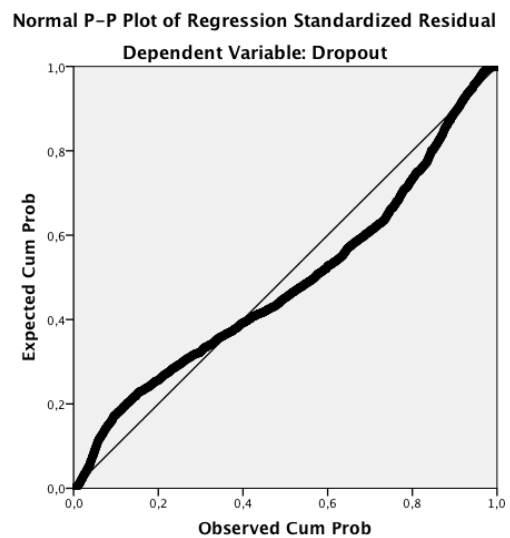


Figure 2. Normal P-Plot of dropout

For dropout, it seems that the data pretty much follows normal distribution, gathered from the visualization in figures 1 and 2, however to be sure Kolmogorov-Smirnov test is executed.

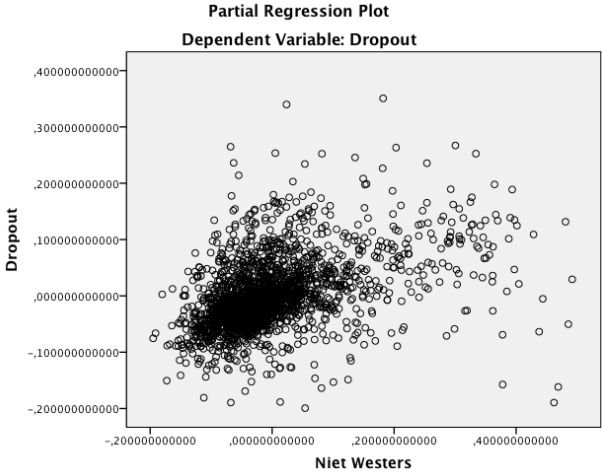
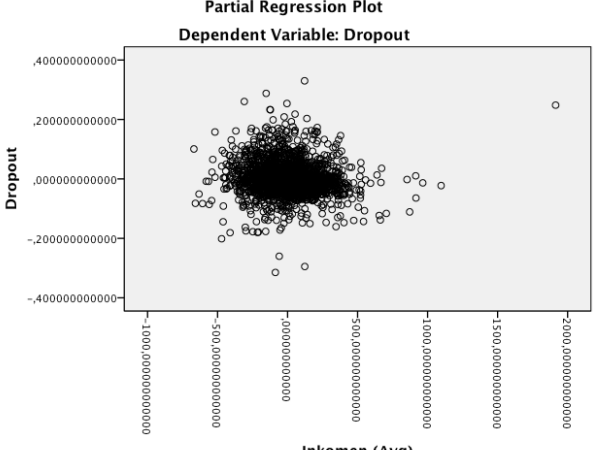
Which leads to the following table:

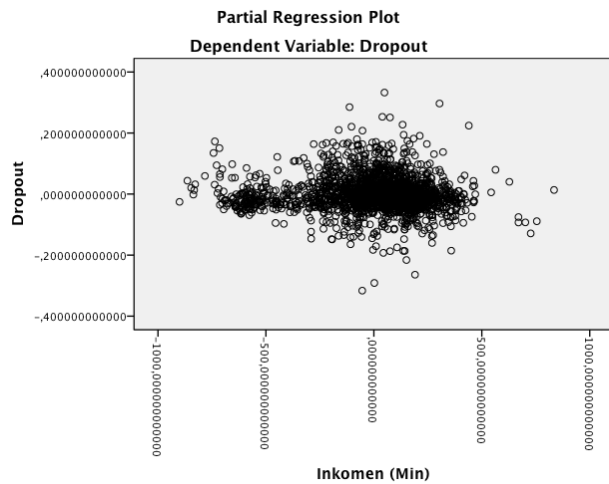
Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Dropout	,210	3738	,000	,691	3738	,000

a. Lilliefors Significance Correction

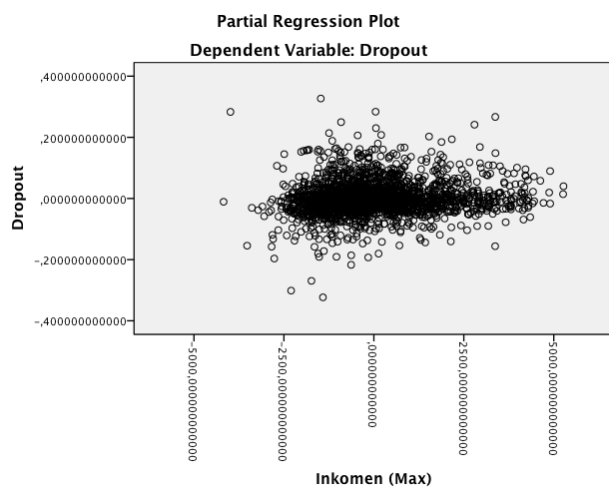
Table 27. K-S test for dropout

Table 23 shows that the percentage of dropout, $D(3738) = 0.21$, $p < .001$, which indicates that dropout is significantly non-normal. Therefore, the model might not be as accurate as needed to draw final conclusions. For now, we just bear in mind that the conclusions are not final. Partial plots were requested, which are scatterplots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors. These scatterplots show the relationships (linear / non-linear) and assess homoscedasticity of the predictors and dropout. Homoscedasticity shows if the variability of a variable is equal across the range of values of a second variable that predicts it.

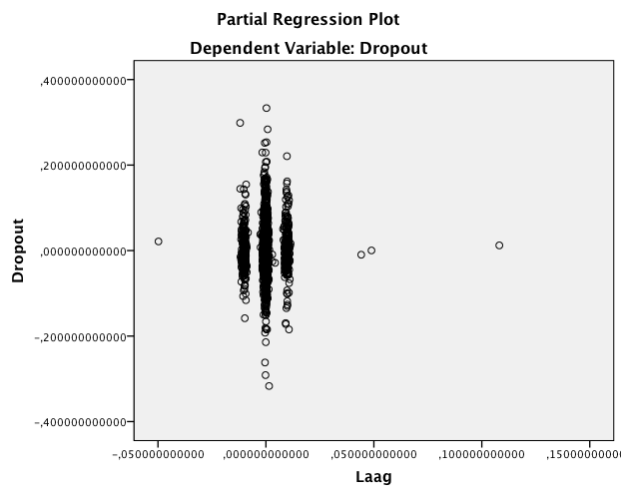
 <p>Partial Regression Plot Dependent Variable: Dropout</p> <p>The plot shows a scatter of points with a slight positive trend. The y-axis is labeled 'Dropout' and ranges from -0.2000000000000000 to 0.4000000000000000. The x-axis is labeled 'Niet Westers' and ranges from -0.2000000000000000 to 0.4000000000000000.</p>	<p>For non-Western ethnicity the partial plot shows a positive relationship to dropout. There are no obvious outliers on this plot, and the cloud of dots is evenly spaced around a line, indicating homoscedasticity.</p>
 <p>Partial Regression Plot Dependent Variable: Dropout</p> <p>The plot shows a scatter of points with a very weak negative trend. The y-axis is labeled 'Dropout' and ranges from -0.4000000000000000 to 0.4000000000000000. The x-axis is labeled 'Inkomen (Avg)' and ranges from 0.0000000000000000 to 2.0000000000000000.</p>	<p>For average income the partial plot shows a very weak negative relationship to dropout. There is one obvious outlier on this plot, and the cloud of dots is evenly spaced around a line, indicating homoscedasticity.</p>



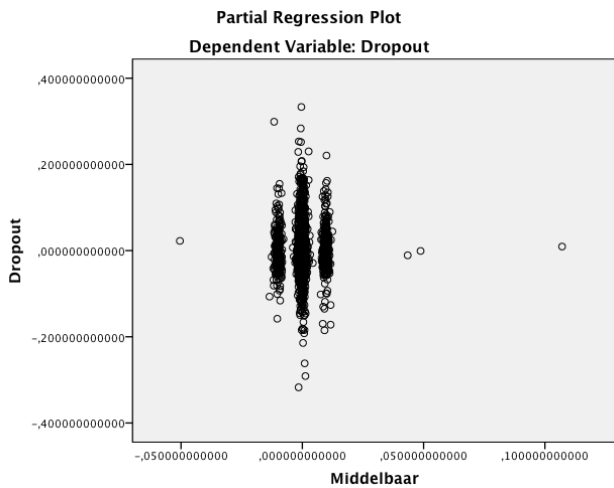
For minimum income the partial plot shows a very weak negative relationship to dropout. There are no obvious outliers on this plot, and the cloud of dots is evenly spaced around a line, indicating homoscedasticity.



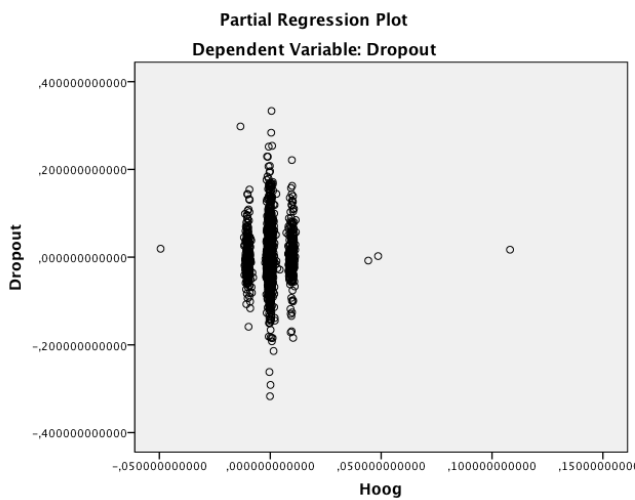
For maximum income the partial plot shows a very weak positive relationship to dropout. There is one obvious outlier on this plot, and the cloud of dots is evenly spaced around a line, indicating homoscedasticity.



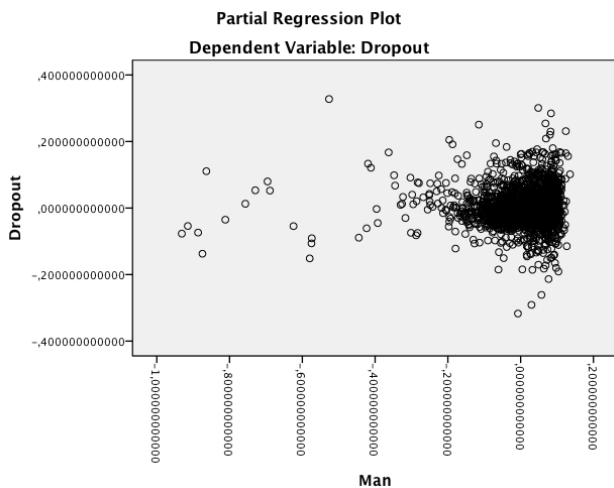
For low education (not in model 4) something interesting is going on, and this indicates that the different levels of education are interrelated, and should be assessed in a factor analysis (see below).



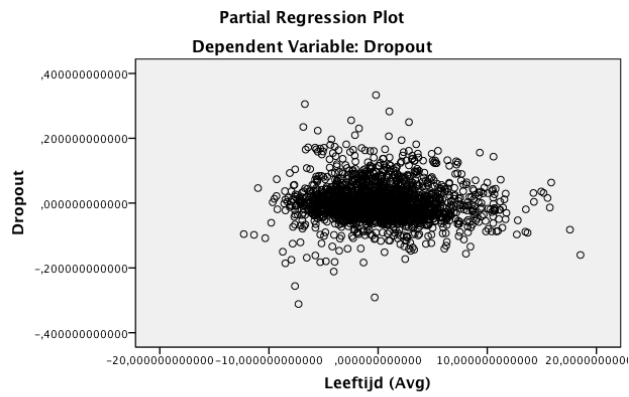
For medium education (not in model 4) something interesting is going on, and this indicates that the different levels of education are interrelated, and should be assessed in a factor analysis (see below).



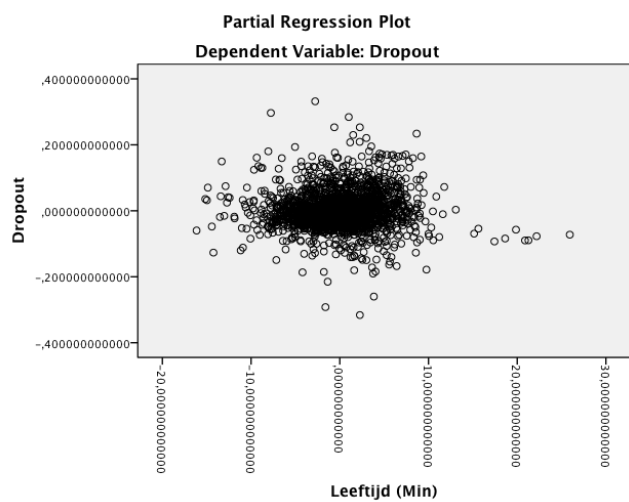
For high education something interesting is going on, and this indicates that the different levels of education are interrelated, and should be assessed in a factor analysis (see below).



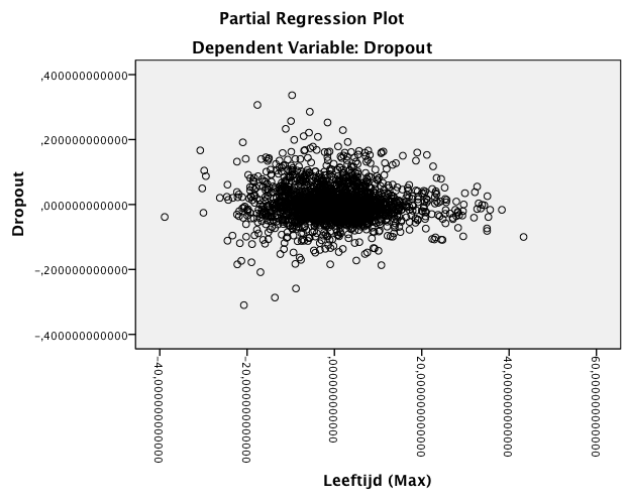
For male gender (not in model 4) the partial plot shows a weak positive relationship to dropout. However, the relationship looks like a funnel, indicating heteroscedasticity (showing greater variance at high levels of male gender).



For average age the partial plot shows a very weak negative relationship to dropout. There is one obvious outlier on this plot, and the cloud of dots is evenly spaced around a line, indicating homoscedasticity.



For minimum age the partial plot shows an ambiguous relationship to dropout. There are some obvious outliers on this plot, and the cloud of dots is evenly spaced around a line, indicating homoscedasticity.



For maximum age the partial plot shows a very weak negative relationship to dropout. There are some obvious outliers on this plot, and the cloud of dots is evenly spaced around a line, indicating homoscedasticity.

4. Multiple regression - output

Descriptive Statistics			
	Mean	Std. Deviation	N
Dropout	,12988739670150	,067133322205874	2181
Westers	,89213760438983	,112728987257569	2181
Niet Westers	,10786239561017	,112728987257569	2181
Inkomen (Avg)	2440,61284980366870	264,638489976578600	2181
Inkomen (Min)	1164,87849610270520	290,499169258922100	2181
Inkomen (Max)	6087,57450710683100	2029,490521001774400	2181
Laag	,47852283529773	,072450305258876	2181
Middelbaar	,34864376539549	,037754259732394	2181
Hoog	,17302892492954	,076889207011452	2181
Man	,90324983634192	,100465495018832	2181
Vrouw	,09662414130202	,100492887936104	2181
Leeftijd (Avg)	28,15695336888475	5,713134360708809	2181
Leeftijd (Min)	7,46125630444750	6,955889068472920	2181
Leeftijd (Max)	79,13617606602476	12,973907401424194	2181

a. Correlations table

Correlations															
		Drop out	Westers	Niet Westers	Inkomen (Avg)	Inkomen (Min)	Inkomen (Max)	Laag	Middelbaar	Hoog	Man	Vrouw	Leeftijd (Avg)	Leeftijd (Min)	Leeftijd (Max)
Pearson Correlation	Dropout	1,000	-,533	,533	-,052	-,123	,209	-,084	-,216	,187	,103	-,104	-,092	-,017	-,066
	Westers	-,533	1,000	.	,037	,081	-,241	,092	,360	-,264	-,172	,172	,036	-,003	,077
	Niet Westers	,533	.	1,000	-,037	-,081	,241	-,092	-,360	,264	,172	-,172	-,036	,003	-,077

	Inkome n (Avg)	-,052	,037	-,037	1,000	,265	,470	-,350	-,102	,384	-,020	,020	,049	,133	-,009
	Inkome n (Min)	-,123	,081	-,081	,265	1,000	-,181	-,058	-,001	,057	,027	-,027	,426	,461	-,314
	Inkome n (Max)	,209	-,241	,241	,470	-,181	1,000	-,235	-,215	,333	,043	-,043	-,256	-,259	,290
	Laag	-,084	,092	-,092	-,350	-,058	-,235	1,000	-,139	-,872	,001	-,001	-,052	-,114	,095
	Middelb aar	-,216	,360	-,360	-,102	-,001	-,215	-,139	1,000	-,358	-,118	,118	,002	,013	-,011
	Hoog	,187	-,264	,264	,384	,057	,333	-,872	-,358	1,000	,057	-,057	,045	,101	-,083
	Man	,103	-,172	,172	-,020	,027	,043	,001	-,118	,057	1,000	- 1,000	,113	,066	-,011
	Vrouw	-,104	,172	-,172	,020	-,027	-,043	-,001	,118	-,057	- 1,000	1,000	-,114	-,066	,011
	Leeftijd (Avg)	-,092	,036	-,036	,049	,426	-,256	-,052	,002	,045	,113	-,114	1,000	,640	-,233
	Leeftijd (Min)	-,017	-,003	,003	,133	,461	-,259	-,114	,013	,101	,066	-,066	,640	1,000	-,603
	Leeftijd (Max)	-,066	,077	-,077	-,009	-,314	,290	,095	-,011	-,083	-,011	,011	-,233	-,603	1,000
Sig. (1- tailed)	Dropout	.	,000	,000	,008	,000	,000	,000	,000	,000	,000	,000	,000	,217	,001
	Westers	,000	.	,000	,042	,000	,000	,000	,000	,000	,000	,000	,047	,438	,000
	Niet Westers	,000	,000	.	,042	,000	,000	,000	,000	,000	,000	,000	,047	,438	,000
	Inkome n (Avg)	,008	,042	,042	.	,000	,000	,000	,000	,000	,176	,176	,011	,000	,339
	Inkome n (Min)	,000	,000	,000	,000	.	,000	,003	,474	,004	,104	,105	,000	,000	,000
	Inkome n (Max)	,000	,000	,000	,000	,000	.	,000	,000	,000	,022	,021	,000	,000	,000
	Laag	,000	,000	,000	,000	,003	,000	.	,000	,000	,483	,483	,008	,000	,000
	Middelb aar	,000	,000	,000	,000	,474	,000	,000	.	,000	,000	,000	,464	,273	,311
	Hoog	,000	,000	,000	,000	,004	,000	,000	,000	.	,004	,004	,017	,000	,000

b. Model summary

Model Summary ^a										
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,558 ^a	,311	,307	,055871311167714	,311	88,948	11	2169	,000	
2	,557 ^b	,311	,308	,055862925458163	,000	,349	1	2169	,555	
3	,557 ^c	,311	,308	,055856838336441	,000	,527	1	2170	,468	
4	,557 ^d	,310	,308	,055855338071588	,000	,883	1	2171	,347	
5	,557 ^e	,311	,307	,055868849986741	,000	,475	2	2170	,622	
6	,558 ^f	,311	,307	,055871679437215	,000	,780	1	2169	,377	1,964
a. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Vrouw, Middelbaar, Leeftijd (Avg), Laag, Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog										
b. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Vrouw, Leeftijd (Avg), Laag, Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog										
c. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Vrouw, Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog										
d. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog										
e. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog, Middelbaar, Laag										
f. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog, Middelbaar, Laag, Man										
g. Dependent Variable: Dropout										

c. ANOVA

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,054	11	,278	88,948	,000 ^b
	Residual	6,771	2169	,003		
	Total	9,825	2180			
2	Regression	3,053	10	,305	97,837	,000 ^c
	Residual	6,772	2170	,003		
	Total	9,825	2180			
3	Regression	3,052	9	,339	108,673	,000 ^d
	Residual	6,773	2171	,003		
	Total	9,825	2180			
4	Regression	3,049	8	,381	122,153	,000 ^e
	Residual	6,776	2172	,003		
	Total	9,825	2180			
5	Regression	3,052	10	,305	97,770	,000 ^f
	Residual	6,773	2170	,003		
	Total	9,825	2180			
6	Regression	3,054	11	,278	88,944	,000 ^g
	Residual	6,771	2169	,003		
	Total	9,825	2180			
a. Dependent Variable: Dropout						
b. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Vrouw, Middelbaar, Leeftijd (Avg), Laag, Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog						
c. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Vrouw, Leeftijd (Avg), Laag, Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog						
d. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Vrouw, Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog						
e. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog						
f. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog, Middelbaar, Laag						
g. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog, Middelbaar, Laag, Man						

d. Model parameters

Coefficients ^a													
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	,055	,207		,265	,791	-,352	,462					
	Niet Westers	,276	,012	,464	22,592	,000	,252	,300	,533	,436	,403	,754	1,326
	Inkomen (Avg)	- 2,732E-5	,000	-,108	-4,545	,000	,000	,000	-,052	-,097	- ,081	,566	1,767
	Inkomen (Min)	- 1,233E-5	,000	-,053	-2,449	,014	,000	,000	-,123	-,053	- ,044	,669	1,494
	Inkomen (Max)	4,369E-6	,000	,132	5,426	,000	,000	,000	,209	,116	,097	,536	1,865
	Laag	,146	,207	,158	,706	,480	-,260	,553	-,084	,015	,013	,006	157,723
	Middelbaar	,123	,209	,069	,590	,555	-,286	,532	-,216	,013	,011	,023	43,275
	Hoog	,191	,208	,219	,921	,357	-,216	,599	,187	,020	,016	,006	178,229
	Vrouw	-,011	,012	-,016	-,899	,369	-,035	,013	-,104	-,019	- ,016	,949	1,054
	Leeftijd (Avg)	-,001	,000	-,065	-2,590	,010	-,001	,000	-,092	-,056	- ,046	,506	1,975
	Leeftijd (Min)	,000	,000	,050	1,658	,097	,000	,001	-,017	,036	,030	,355	2,818
Leeftijd (Max)	,000	,000	-,067	-2,795	,005	-,001	,000	-,066	-,060	- ,050	,546	1,830	
2	(Constant)	,176	,027		6,555	,000	,124	,229					
	Niet Westers	,276	,012	,463	22,589	,000	,252	,300	,533	,436	,403	,755	1,325
	Inkomen (Avg)	- 2,731E-	,000	-,108	-4,543	,000	,000	,000	-,052	-,097	- ,081	,566	1,767

		5											
	Inkomen (Min)	- 1,226E-5	,000	-,053	-2,436	,015	,000	,000	-,123	-,052	- ,043	,670	1,493
	Inkomen (Max)	4,387E-6	,000	,133	5,453	,000	,000	,000	,209	,116	,097	,537	1,862
	Laag	,026	,036	,028	,726	,468	-,044	,096	-,084	,016	,013	,216	4,634
	Hoog	,070	,036	,081	1,983	,047	,001	,140	,187	,043	,035	,192	5,216
	Vrouw	-,011	,012	-,016	-,897	,370	-,035	,013	-,104	-,019	- ,016	,949	1,053
	Leeftijd (Avg)	-,001	,000	-,065	-2,616	,009	-,001	,000	-,092	-,056	- ,047	,507	1,971
	Leeftijd (Min)	,000	,000	,050	1,676	,094	,000	,001	-,017	,036	,030	,355	2,816
	Leeftijd (Max)	,000	,000	-,067	-2,787	,005	-,001	,000	-,066	-,060	- ,050	,546	1,830
3	(Constant)	,192	,016		11,907	,000	,160	,224					
	Niet Westers	,278	,012	,467	23,592	,000	,255	,301	,533	,452	,420	,810	1,235
	Inkomen (Avg)	- 2,742E-5	,000	-,108	-4,565	,000	,000	,000	-,052	-,098	- ,081	,566	1,766
	Inkomen (Min)	- 1,209E-5	,000	-,052	-2,405	,016	,000	,000	-,123	-,052	- ,043	,671	1,490
	Inkomen (Max)	4,423E-6	,000	,134	5,509	,000	,000	,000	,209	,117	,098	,539	1,855
	Hoog	,048	,018	,055	2,686	,007	,013	,083	,187	,058	,048	,751	1,331
	Vrouw	-,011	,012	-,017	-,940	,347	-,035	,012	-,104	-,020	- ,017	,952	1,050
	Leeftijd (Avg)	-,001	,000	-,065	-2,609	,009	-,001	,000	-,092	-,056	- ,046	,507	1,971
	Leeftijd (Min)	,000	,000	,050	1,660	,097	,000	,001	-,017	,036	,030	,355	2,814
	Leeftijd	,000	,000	-,067	-2,768	,006	-,001	,000	-,066	-,059	-	,547	1,828

	(Max)										,049			
4	(Constant)	,190	,016		11,884	,000	,159	,222						
	Niet Westers	,280	,012	,470	23,998	,000	,257	,303	,533	,458	,428	,828	1,208	
	Inkomen (Avg)	- 2,767E-5	,000	-,109	-4,611	,000	,000	,000	,000	-,052	-,098	- ,082	,567	1,762
	Inkomen (Min)	- 1,204E-5	,000	-,052	-2,396	,017	,000	,000	,000	-,123	-,051	- ,043	,671	1,490
	Inkomen (Max)	4,456E-6	,000	,135	5,556	,000	,000	,000	,000	,209	,118	,099	,540	1,851
	Hoog	,048	,018	,055	2,696	,007	,013	,084	,187	,058	,048	,048	,751	1,331
	Leeftijd (Avg)	-,001	,000	-,063	-2,533	,011	-,001	,000	,000	-,092	-,054	- ,045	,512	1,954
	Leeftijd (Min)	,000	,000	,050	1,670	,095	,000	,001	,001	-,017	,036	,030	,355	2,814
	Leeftijd (Max)	,000	,000	-,066	-2,748	,006	-,001	,000	,000	-,066	-,059	- ,049	,547	1,828
5	(Constant)	,053	,207		,255	,799	-,354	,460						
	Niet Westers	,278	,012	,466	22,897	,000	,254	,301	,533	,441	,408	,766	1,305	
	Inkomen (Avg)	- 2,755E-5	,000	-,109	-4,588	,000	,000	,000	,000	-,052	-,098	- ,082	,567	1,764
	Inkomen (Min)	- 1,230E-5	,000	-,053	-2,443	,015	,000	,000	,000	-,123	-,052	- ,044	,669	1,494
	Inkomen (Max)	4,398E-6	,000	,133	5,468	,000	,000	,000	,000	,209	,117	,097	,537	1,862
	Laag	,148	,207	,159	,711	,477	-,259	,554	-,084	,015	,013	,006	157,718	
	Middelbaar	,122	,208	,069	,587	,557	-,287	,531	-,216	,013	,010	,023	43,274	
	Hoog	,192	,208	,220	,926	,355	-,215	,600	,187	,020	,016	,006	178,224	
	Leeftijd (Avg)	-,001	,000	-,063	-2,519	,012	-,001	,000	,000	-,092	-,054	- ,045	,511	1,958

	Leeftijd (Min)	,000	,000	,050	1,669	,095	,000	,001	-,017	,036	,030	,355	2,817
	Leeftijd (Max)	,000	,000	-,067	-2,777	,006	-,001	,000	-,066	-,060	- ,050	,547	1,830
6	(Constant)	,044	,208		,213	,832	-,363	,451					
	Niet Westers	,276	,012	,464	22,596	,000	,252	,300	,533	,437	,403	,754	1,326
	Inkomen (Avg)	- 2,732E-5	,000	-,108	-4,546	,000	,000	,000	-,052	-,097	- ,081	,566	1,767
	Inkomen (Min)	- 1,233E-5	,000	-,053	-2,449	,014	,000	,000	-,123	-,053	- ,044	,669	1,494
	Inkomen (Max)	4,370E-6	,000	,132	5,427	,000	,000	,000	,209	,116	,097	,536	1,865
	Laag	,146	,207	,158	,706	,480	-,260	,553	-,084	,015	,013	,006	157,723
	Middelbaar	,123	,209	,069	,590	,555	-,286	,532	-,216	,013	,011	,023	43,275
	Hoog	,191	,208	,219	,921	,357	-,216	,599	,187	,020	,016	,006	178,229
	Leeftijd (Avg)	-,001	,000	-,065	-2,589	,010	-,001	,000	-,092	-,056	- ,046	,506	1,975
	Leeftijd (Min)	,000	,000	,050	1,658	,097	,000	,001	-,017	,036	,030	,355	2,818
	Leeftijd (Max)	,000	,000	-,067	-2,795	,005	-,001	,000	-,066	-,060	- ,050	,546	1,830
	Man	,011	,012	,016	,883	,377	-,013	,035	,103	,019	,016	,949	1,053
a. Dependent Variable: Dropout													

e. Excluded variables

Model		Excluded Variables ^a						
		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Westers	. ^b	.	.	.	,000	.	,000
	Man	-4,675 ^b	-2,024	,043	-,043	5,949E-5	16809,190	5,947E-5
2	Westers	. ^c	.	.	.	,000	.	,000
	Man	-4,659 ^c	-2,018	,044	-,043	5,950E-5	16807,092	5,948E-5
	Middelbaar	,069 ^c	,590	,555	,013	,023	43,275	,006
3	Westers	. ^d	.	.	.	,000	.	,000
	Man	-4,640 ^d	-2,010	,045	-,043	5,951E-5	16805,136	5,949E-5
	Middelbaar	-,012 ^d	-,614	,539	-,013	,786	1,272	,355
	Laag	,028 ^d	,726	,468	,016	,216	4,634	,192
4	Westers	. ^e	.	.	.	,000	.	,000
	Man	,017 ^e	,924	,356	,020	,953	1,050	,355
	Middelbaar	-,013 ^e	-,666	,505	-,014	,789	1,267	,355
	Laag	,030 ^e	,778	,437	,017	,216	4,619	,192
	Vrouw	-,017 ^e	-,940	,347	-,020	,952	1,050	,355
5	Westers	. ^f	.	.	.	,000	.	,000
	Man	,016 ^f	,883	,377	,019	,949	1,053	,006
	Vrouw	-,016 ^f	-,899	,369	-,019	,949	1,054	,006
6	Westers	. ^g	.	.	.	,000	.	,000
	Vrouw	-4,692 ^g	-2,031	,042	-,044	5,947E-5	16813,957	5,947E-5
a. Dependent Variable: Dropout								
b. Predictors in the Model: (Constant), Leeftijd (Max), Inkomen (Avg), Vrouw, Middelbaar, Leeftijd (Avg), Laag, Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog								
c. Predictors in the Model: (Constant), Leeftijd (Max), Inkomen (Avg), Vrouw, Leeftijd (Avg), Laag, Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog								
d. Predictors in the Model: (Constant), Leeftijd (Max), Inkomen (Avg), Vrouw, Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog								
e. Predictors in the Model: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog								
f. Predictors in the Model: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog,								

Middelbaar, Laag

g. Predictors in the Model: (Constant), Leeftijd (Max), Inkomen (Avg), Leeftijd (Avg), Niet Westers, Inkomen (Min), Inkomen (Max), Leeftijd (Min), Hoog, Middelbaar, Laag, Man

f. Casewise diagnostics

Casewise Diagnostics ^a				
Case Number	Std. Residual	Dropout	Predicted Value	Residual
12	2,659	,240121580547	,09153554954827	,148586030998845
13	-4,686	,000000000000	,26178990818664	-,261789908186635
24	5,378	,380434782609	,07998098362783	,300453798980861
96	-2,560	,000000000000	,14302237676780	-,143022376767805
98	3,158	,393335962145	,21691200852100	,176423953624109
151	-3,302	,142156862745	,32665935655904	-,184502493813938
153	5,076	,414048059150	,13043975875897	,283608300390755
192	2,596	,278012684989	,13297399308618	,145038691903247
242	3,205	,318932655654	,13987710730890	,179055548345483
309	2,392	,247524752475	,11387038166577	,133654370809483
326	-2,453	,007936507937	,14499467179731	-,137058163860804
360	-2,055	,050847457627	,16563592198550	-,114788464358383
367	-2,530	,018867924528	,16023240866804	-,141364484139736
375	-2,041	,090277777778	,20432818946151	-,114050411683730
382	2,502	,251184834123	,11139011325750	,139794720865725
383	2,099	,294498381877	,17724351740398	,117254864473047
389	2,981	,346726190476	,18015802721628	,166568163259907
425	-2,407	,000000000000	,13449325763463	-,134493257634635
427	4,510	,412408759124	,16041155741657	,251997201707517
430	2,504	,261363636364	,12148535799111	,139878278372521
456	4,101	,381909547739	,15275667883092	,229152868907775
490	2,944	,282024793388	,11752620986545	,164498583522984
526	2,246	,250830564784	,12533119418595	,125499370598106
531	3,505	,328269484808	,13242708220298	,195842402605470

538	-2,243	,000000000000	,12530913315218	-,125309133152183
566	2,036	,226527570790	,11274569375874	,113781877031131
637	-2,487	,008849557522	,14777528009565	-,138925722573527
641	2,279	,282226562500	,15488360386543	,127342958634574
684	2,899	,295882053889	,13391601185518	,161966042033993
695	2,148	,257653061224	,13761764934998	,120035411874508
698	3,464	,290909090909	,09736271029884	,193546380610252
817	2,991	,376344086022	,20923286586096	,167111220160542
819	2,889	,283146067416	,12175393384407	,161392133571662
903	2,030	,293752980448	,18034368720535	,113409293242910
916	-2,268	,067901234568	,19461935909492	-,126718124527015
933	2,137	,237931034483	,11853046669382	,119400567788940
942	-2,612	,000000000000	,14591230102632	-,145912301026323
981	-3,331	,000000000000	,18611241575715	-,186112415757154
982	2,992	,274193548387	,10704302805819	,167150520328908
985	2,268	,232067510549	,10534491176121	,126722598787311
1007	-3,438	,000000000000	,19208706006175	-,192087060061748
1011	-2,807	,000000000000	,15684424858888	-,156844248588875
1018	-2,362	,000000000000	,13194980260821	-,131949802608207
1027	2,873	,286123032904	,12558925082938	,160533782074769
1038	3,706	,393422655298	,18635636929053	,207066286007884
1047	2,306	,278709677419	,14987018461081	,128839492808544
1051	2,389	,254748603352	,12129388886524	,133454714486713
1064	-2,265	,151898734177	,27844728645521	-,126548552277997
1098	-3,835	,000000000000	,21428407751505	-,214284077515048
1102	2,869	,287949921753	,12766137365563	,160288548097105
1118	3,064	,306990881459	,13577838639740	,171212495061564
1138	-3,103	,076190476190	,24954713369671	-,173356657506235
1177	3,294	,433161216294	,24910382963640	,184057386657350
1212	2,963	,374868004224	,20934558303595	,165522421187920
1218	2,225	,265317594154	,14102339811767	,124294196036344
1319	2,789	,266247379455	,11042829836597	,155819081088955
1379	-3,302	,132739420935	,31723264147377	-,184493220538361
1418	2,421	,243397573162	,10815689231945	,135240680842580

1451	2,536	,298623063683	,15692453857132	,141698525111980
1452	2,327	,290919952210	,16088252312531	,130037429084969
1470	5,965	,438356164384	,10509449845455	,333261665929011
1515	-2,078	,000000000000	,11612465660176	-,116124656601757
1545	2,245	,217488789238	,09207333514647	,125415454091195
1564	-3,276	,000000000000	,18306149121138	-,183061491211375
1574	2,406	,266283524904	,13186403084301	,134419494061210
1620	2,401	,246346555324	,11218133709313	,134165218230461
1635	-3,056	,000000000000	,17074446177430	-,170744461774302
1647	2,406	,248697916667	,11426897902089	,134428937645776
1666	2,785	,267683772538	,11207737156454	,155606400973597
1679	2,065	,232227488152	,11684424894761	,115383239204054
1683	-2,060	,129169104740	,24428074437632	-,115111639636709
1691	2,122	,288224956063	,16963900246544	,118585953597826
1720	2,616	,308087291399	,16195389569199	,146133395707236
1730	2,844	,291933418694	,13301936319237	,158914055501608
1741	-2,969	,000000000000	,16589414848703	-,165894148487031
1790	-2,435	,000000000000	,13604354336886	-,136043543368858
1815	3,724	,403697996918	,19562824691148	,208069750006857
1928	2,783	,362244897959	,20674674129649	,155498156662693
1964	2,050	,271645736946	,15711201088323	,114533726063238
2063	2,058	,248664400194	,13365784803519	,115006552159080
2098	2,461	,278481012658	,14100760351325	,137473409144975
2109	2,114	,283323716099	,16523163437627	,118092081722985
2124	2,270	,280000000000	,15317673643315	,126823263566852
2133	2,874	,300375469337	,13982290935729	,160552559979379
2136	2,867	,402286902287	,24211884685977	,160168055427133
2138	2,466	,261417322835	,12361907867096	,137798244163683
2187	-2,195	,000000000000	,12264909434340	-,122649094343396
2192	-2,702	,000000000000	,15099130409134	-,150991304091339
2268	-2,682	,000000000000	,14985709209968	-,149857092099683
2315	2,400	,250618301731	,11651233500613	,134105966725112
2333	2,709	,309045226131	,15768862287085	,151356603259805
2367	2,763	,268011527378	,11363005657719	,154381470800332

2394	-2,039	,002192146397	,11610140412141	-,113909257724151
2398	2,080	,236024844720	,11983168177679	,116193162943706
2401	2,534	,238619309655	,09706184343996	,141557466214871
2416	-2,045	,178010471204	,29229543745480	-,114284966250616
2463	2,918	,343360995851	,18034282549809	,163018170352529
2555	2,791	,329787234043	,17382910672252	,155958127320031
2566	2,222	,267489711934	,14331819768645	,124171514247710
2577	2,914	,272160664820	,10932505410594	,162835610714001
2581	2,571	,314691151920	,17106744830681	,143623703613052
2610	4,112	,333333333333	,10357284701902	,229760486314312
2680	2,361	,351624351624	,21971391074553	,131910440878821
2715	-5,675	,024193548387	,34128342815818	-,317089879771080
2730	-2,299	,057142857143	,18556474476565	-,128421887622795
2753	-5,214	,000000000000	,29130195442778	-,291301954427782
2810	2,984	,282186948854	,11545841279489	,166728536058721
3018	2,453	,306194690265	,16915872588762	,137035964377865
3036	-2,147	,009569377990	,12951817004422	-,119948792053789
3039	-2,090	,000000000000	,11679117487558	-,116791174875577
3040	-2,597	,022435897436	,16754506036468	-,145109162928783
3098	3,926	,377450980392	,15808748851216	,219363491880000
3194	3,023	,269146608315	,10026459581357	,168882012501523
3373	-2,527	,127145085803	,26830870150177	-,141163615698337
3375	2,455	,238826815642	,10167699764343	,137149817999027
3452	2,366	,259367681499	,12717216554062	,132195515958206
3501	-2,287	,055555555556	,18333816399320	-,127782608437643
3560	2,069	,245084269663	,12947192561127	,115612344051654
3608	2,654	,245098039216	,09682545722420	,148272581991485
3655	-2,348	,000000000000	,13116061078048	-,131160610780478
3661	-2,007	,000000000000	,11213278156043	-,112132781560435
3674	4,538	,405844155844	,15228141918306	,253562736661098
3685	2,342	,247524752475	,11669006995113	,130834682524122
3691	-3,238	,000000000000	,18092729119786	-,180927291197860
3711	-2,148	,063414634146	,18343623537074	-,120021601224400
3714	2,709	,311572700297	,16024020642798	,151332493868757

3715	2,926	,275426405559	,11192810462447	,163498300934597
3727	-2,684	,088000000000	,23796033873212	-,149960338732120
a. Dependent Variable: Dropout				

5. Factor analysis

Factor analysis (or in this case principal components analysis) is used to understand the structure of a set of variables, in this case the variables regarding education. In addition, this factor analysis combines the education variables that have proven collinear in the multiple regression discussed above. In this case, I chose to do Promax factor rotation, which is an oblique rotation method used for large datasets that allows the factors to correlate (as different education levels probably do).

A principal components analysis (PCA) was conducted on the three variables (low, medium, and high education), with oblique rotation (Promax). The Kaiser-Meyer-Olkin measure did not verify the sampling adequacy to the analysis, $KMO = .236$, which is odd, however, Bartlett's test of sphericity $\chi^2 (3) = 11311.372$, $p < .001$, indicated that the relationship between variables was sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Two components had eigenvalues over Kaiser's criterion of 1 and in combination explained 99.911% of the variance. The scree plot was slightly ambiguous and showed inflections that would justify retaining component 1. Given the large sample size, and the convergence of the scree plot and Kaiser's criterion on two components, this is the number of components that were retained for the final analysis. Table 25 shows the factor loadings after rotation. The variables that cluster on the same components suggest that component 1 represents low and high education, and component 2 represent medium education.

	Component	
	1	2
Low education	,534	-,205
Medium education	-,005	,885
High education	-,501	-,240
Eigenvalues	1.901	1.096
% of the variance	63.365	36.546
Extraction Method: PCA. Rotation Method: Promax with Kaiser Normalization.		
Note: Factor loadings above .40 appear in bold.		

Table 28. Summary of PCA results for education

As oblique rotation was used, tables of the structure and pattern matrices are given. The structure matrix shows the correlation coefficient between each variable and factor. The pattern matrix shows the regression coefficients for each variable on each factor.

	Component	
	1	2
Low education	,971	
High education	-,964	
Medium education		1,000
Extraction Method: PCA. Rotation Method: Promax with Kaiser Normalization.		

Table 29. Structure mix

	Component	
	1	2
Low education	1,000	
High education	-,932	
Medium education		1,002
Extraction Method: PCA. Rotation Method: Promax with Kaiser Normalization. ^a		
a. Rotation converged in 3 iterations.		

Table 30. Pattern mix

Both structure matrix and pattern matrix show similar values of the variables on the components. However, it is strange that medium education loads highly onto a different component than low and high education. To assess whether the PCA actually led to a better way of explaining dropout, simple regression is conducted with component 1.

Simple regression with education

This section will look into the simple regression method, trying to assess if education after PCA can explain the variance in dropout better than before. Table 28 provides the values of R and R² for the model that has been derived. For these data, R has a value of .124 and because there is only one predictor, this value represents the simple correlation between education (low and high) and record sales. The value of R² is .015, which tells us that education can account for only 1.5% of the variation in dropout. This means that 98.5% of the variation in dropout cannot be explained by education alone. Therefore there must be other variables that have an influence also (see above).

	R	R ²	Adjusted R ²	SE of the Estimate	Change Statistics					Durbin-Watson
					R ² Change	F Change	df 1	df2	Sig. F Change	
1	,124 ^a	,015	,015	,06662701 5419432	,015	34,258	1	21 79	,000	1,920
a. Predictors: (Constant), Education (high and low).										
b. Dependent Variable: Dropout										

Table 31. Model summary

For these data F is 34.258, which is significant at $p < .001$. This result tells us that there is less than 0.1% chance that an F-ratio this large would happen if the null hypothesis were true. Therefore, we can conclude that this regression model results in significantly better prediction of dropout than if we used the mean value of dropout. In short, the regression model overall predicts dropout significantly well. From table 29 we can say that b_0 is .130 and this can be interpreted as meaning that when education is 0, the model predicts that 0.13% dropout will occur. We can also read off the value of b_1 from the table, and this value represents the gradient of the regression line. It is -.008. If our predictor variable increased with one unit, then our model predicts a increase in dropout of -.008.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,130	,001		91,043	,000
	Education 1	-,008	,001	-,124	-5,853	,000
a. Dependent Variable: Dropout						

Table 32. Coefficients

The t-test tells us whether the b value is different from 0. If the observed significance is less than .05 the results reflect a genuine effect. For these two values the probabilities are .000 and so we can say that the probabilities of these t-values or larger occurring if the values of b in the population were 0 is less than .001. Therefore, the bs are different from 0 and we can conclude that education makes a significant contribution ($p < .001$) to predicting dropout.

However, PCA did not succeed in creating a better latent variable that includes all aspects of the education variables. Therefore, a new multiple regression is executed in the next section.

6. Multiple regression new model

In this section a new multiple regression model is constructed using only high education and the other variables already present in model 4. The outcomes of this model are briefly discussed in this section on aspects where the new model differs substantially from model 4.

	R	R ²	Adjusted R ²	SE of the Estimate	Change Statistics					Durbin-Watson
					R ² Change	F Change	df 1	df2	Sig. F Change	
1	,557 ^a	,310	,308	,05575760 3456542	,310	123,145	8	21 92	,000	1,960
a. Predictors: (Constant), Leeftijd (Max), Inkomen (Avg), Niet Westers, Leeftijd (Avg), Hoog, Inkomen (Min), Inkomen (Max), Leeftijd (Min)										
b. Dependent Variable: Dropout										

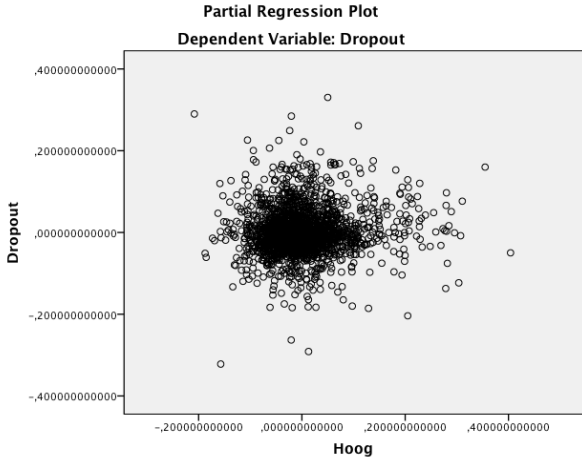
Table 33. Model summary

The model summary shows that the model has slightly improved, as the F change is a little bit higher. All other values are similar to the outcomes given above. The coefficient summary (table 31) shows that the b and standardized β values for all variables have not changed very much.

Model		Unstandardized Coefficients		Standardized Coefficients
		B	Std. Error	Beta
1	(Constant)	,190	,016	
	Hoog	,047	,018	,054
	Niet Westers	,280	,012	,470
	Inkomen (Avg)	-2,810E-5	,000	-,111
	Inkomen (Min)	-1,173E-5	,000	-,051
	Inkomen (Max)	4,538E-6	,000	,137
	Leeftijd (Avg)	-,001	,000	-,062
	Leeftijd (Min)	,000	,000	,050
	Leeftijd (Max)	,000	,000	-,066
a. Dependent Variable: Dropout				

Table 34. Coefficients

The casewise diagnostics give the same cases that have standardized residuals larger than +/- 2 as in model 4. The VIF and tolerance are both relatively high, and thus collinearity should be assessed further. The histogram and normal P-P plots of dropout show the same results, as do the partial plots of the variables non-Western ethnicity, average income, minimum income, maximum income, minimum age, and maximum age. However, the partial plot for high education has changed considerably, as shown below.



For high education the partial plot shows a weak positive relationship to dropout. There are no obvious outliers on this plot, and the cloud of dots is evenly spaced around a line, indicating homoscedasticity. However, also this partial plot indicates the need for further investigation of the (new) model.

The outcomes of this new model underpin the conclusion made after the first round of multiple regression analyses.

8. Encore - output

Descriptive Statistics			
	Mean	Std. Deviation	N
Dropout	,189954	,1751696	3737
Ethnicityindex	,785933	,2062363	3737
Genderindex	,826684	,1596699	3737

Correlations				
		Dropout	Ethnicityindex	Genderindex
Pearson Correlation	Dropout	1,000	-,253	,210
	Ethnicityindex	-,253	1,000	-,210
	Genderindex	,210	-,210	1,000
Sig. (1-tailed)	Dropout	.	,000	,000
	Ethnicityindex	,000	.	,000
	Genderindex	,000	,000	.
N	Dropout	3737	3737	3737
	Ethnicityindex	3737	3737	3737
	Genderindex	3737	3737	3737

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	Genderindex, Ethnicityindex ^b	.	Enter
a. Dependent Variable: Dropout			
b. All requested variables entered.			

Model Summary ^b										
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	

1	,299 ^a	,090	,089	,1671763	,090	183,903	2	3734	,000	1,785
a. Predictors: (Constant), Genderindex, Ethnicityindex										
b. Dependent Variable: Dropout										

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10,279	2	5,140	183,903	,000 ^b
	Residual	104,357	3734	,028		
	Total	114,637	3736			
a. Dependent Variable: Dropout						
b. Predictors: (Constant), Genderindex, Ethnicityindex						

Coefficients ^a													
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	,188	,020		9,434	,000	,149	,227					
	Ethnicityindex	-,186	,014	-,219	-	,000	-,212	-,159	-,253	-,219	-	,956	1,046
	Genderindex	,179	,018	,164	10,243	,000	,145	,214	,210	,165	,160	,956	1,046
a. Dependent Variable: Dropout													

Collinearity Diagnostics ^a							
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	Ethnicityindex	Genderindex	
1	1	2,926	1,000	,00	,01	,00	
	2	,062	6,878	,01	,59	,22	
	3	,012	15,465	,99	,41	,78	
a. Dependent Variable: Dropout							

Casewise Diagnostics ^a				
Case Number	Std. Residual	Dropout	Predicted Value	Residual
12	-2,130	,0000	,356086	-,3560864
42	3,842	1,0000	,357685	,6423153
61	2,739	,6129	,155081	,4578192
67	3,958	1,0000	,338356	,6616443
70	4,812	1,0000	,195506	,8044936
80	2,080	,5581	,210449	,3476507
99	2,184	,6167	,251646	,3650541
119	2,223	,5600	,188295	,3717048
127	3,868	,9677	,321145	,6465546
201	4,800	1,0000	,197480	,8025198
255	4,897	1,0000	,181270	,8187302
271	2,571	,6226	,192755	,4298446
357	4,324	1,0000	,277107	,7228926
369	2,586	,6222	,189819	,4323808
397	4,897	1,0000	,181270	,8187302
436	2,075	,5714	,224481	,3469186
441	2,798	,6222	,154429	,4677710
479	4,608	1,0000	,229592	,7704075
492	4,568	1,0000	,236368	,7636324
495	2,268	,5455	,166265	,3792354
503	4,634	1,0000	,225338	,7746618
522	2,383	,6341	,235719	,3983811
532	2,233	,6129	,239666	,3732341
542	3,669	,8214	,207996	,6134040
567	2,810	,6829	,213163	,4697372
574	4,224	1,0000	,293824	,7061755
612	2,009	,7027	,366903	,3357968
616	2,725	,5763	,120766	,4555343
642	3,926	,9063	,249888	,6564120
670	2,443	,5000	,091544	,4084563

678	5,281	1,0000	,117090	,8829101
681	2,904	,6667	,181270	,4854302
699	2,101	,6652	,313933	,3512672
706	2,837	,6949	,220597	,4743030
733	4,135	1,0000	,308804	,6911955
735	2,008	,4833	,147579	,3357213
744	4,673	1,0000	,218776	,7812244
747	2,758	,6400	,178847	,4611534
757	2,625	,5128	,073957	,4388428
784	2,421	,6000	,195283	,4047166
820	5,025	1,0000	,159885	,8401147
829	4,656	,9286	,150289	,7783109
850	2,787	,6257	,159734	,4659656
875	4,547	,9474	,187180	,7602200
904	4,707	1,0000	,213126	,7868744
922	4,947	1,0000	,172976	,8270235
939	3,864	1,0000	,353968	,6460324
1012	2,515	,6350	,214571	,4204287
1014	4,503	1,0000	,247286	,7527140
1027	2,172	,4103	,047179	,3631214
1034	3,636	,7632	,155377	,6078234
1041	5,540	1,0000	,073914	,9260865
1078	5,173	1,0000	,135239	,8647614
1081	4,781	1,0000	,200745	,7992553
1111	4,914	1,0000	,178508	,8214919
1119	3,956	1,0000	,338616	,6613843
1129	4,812	1,0000	,195539	,8044608
1145	4,128	1,0000	,309930	,6900697
1218	4,853	1,0000	,188704	,8112959
1222	4,300	1,0000	,281107	,7188934
1293	4,574	1,0000	,235354	,7646460
1327	4,730	1,0000	,209260	,7907402
1359	5,573	1,0000	,068338	,9316622
1414	4,179	1,0000	,301370	,6986298

1423	2,579	,7417	,310626	,4310741
1454	2,288	,6393	,256850	,3824497
1469	2,178	,4384	,074236	,3641636
1494	4,383	1,0000	,267284	,7327158
1502	2,004	,5612	,226242	,3349582
1504	2,364	,5833	,188035	,3952650
1506	4,186	1,0000	,300218	,6997821
1558	4,030	1,0000	,326238	,6737622
1568	2,273	,5862	,206249	,3799511
1575	4,897	1,0000	,181270	,8187302
1624	4,880	1,0000	,184132	,8158680
1649	4,210	1,0000	,296115	,7038852
1656	2,768	,6579	,195098	,4628025
1772	4,835	1,0000	,191734	,8082660
1788	2,371	,6214	,225095	,3963052
1805	4,855	1,0000	,188332	,8116676
1817	2,170	,4909	,128166	,3627340
1827	2,037	,6182	,277729	,3404707
1844	4,842	1,0000	,190600	,8094002
1849	4,786	1,0000	,199855	,8001445
1891	4,897	1,0000	,181270	,8187302
1914	2,618	,4762	,038450	,4377501
1915	4,735	1,0000	,208493	,7915068
1917	2,370	,6302	,234072	,3961284
1932	2,065	,5500	,204762	,3452379
1962	2,160	,4932	,132126	,3610736
1992	2,001	,4169	,082451	,3344489
1993	2,167	,4000	,037708	,3622919
2093	4,865	1,0000	,186622	,8133775
2094	2,533	,6087	,185321	,4233785
2103	2,560	,6214	,193439	,4279608
2125	4,692	1,0000	,215616	,7843839
2136	5,326	1,0000	,109625	,8903754
2141	4,676	1,0000	,218330	,7816704

2172	2,444	,5238	,115284	,4085156
2189	3,554	,8125	,218404	,5940961
2210	3,814	,9189	,281372	,6375279
2220	2,450	,6071	,197588	,4095120
2320	2,357	,6202	,226101	,3940995
2385	4,595	1,0000	,231860	,7681401
2428	3,400	,7778	,209483	,5683172
2450	2,373	,6786	,281855	,3967447
2578	4,752	1,0000	,205580	,7944202
2641	3,933	1,0000	,342570	,6574304
2649	4,482	1,0000	,250654	,7493456
2650	4,707	1,0000	,213088	,7869116
2654	2,650	,6309	,187846	,4430536
2688	2,655	,6400	,196160	,4438400
2692	4,669	1,0000	,219445	,7805553
2772	4,788	1,0000	,199616	,8003837
2778	4,781	1,0000	,200748	,7992524
2779	4,425	1,0000	,260233	,7397671
2807	4,067	1,0000	,320030	,6799698
2810	3,262	,6667	,121440	,5452595
2814	2,343	,5921	,200450	,3916498
2829	5,141	1,0000	,140516	,8594845
2836	2,300	,5931	,208558	,3845416
2904	2,710	,6415	,188407	,4530933
2916	2,447	,5000	,090919	,4090806
2917	2,228	,5714	,198963	,3724367
2918	2,592	,5000	,066702	,4332977
2920	2,582	,5000	,068338	,4316622
2931	2,577	,5000	,069267	,4307329
2936	2,582	,5000	,068375	,4316250
2959	2,588	,5000	,067409	,4325915
2963	2,555	,5000	,072798	,4272016
2965	2,584	,5000	,068078	,4319224
2971	2,580	,5000	,068617	,4313834

2993	2,487	,6182	,202415	,4157847
3018	2,012	,5323	,195915	,3363847
3020	4,578	1,0000	,234695	,7653053
3033	4,910	1,0000	,179102	,8208979
3086	2,508	,6250	,205766	,4192343
3087	2,680	,6456	,197629	,4479713
3093	4,600	1,0000	,230968	,7690322
3100	4,712	1,0000	,212308	,7876922
3101	4,684	1,0000	,217029	,7829714
3143	4,716	1,0000	,211564	,7884356
3150	2,563	,6308	,202383	,4284169
3169	4,765	1,0000	,203387	,7966133
3191	4,961	1,0000	,170569	,8294309
3239	3,218	,6000	,061944	,5380562
3254	4,862	1,0000	,187150	,8128504
3319	3,037	,7021	,194454	,5076463
3359	2,287	,5641	,181797	,3823033
3475	4,052	1,0000	,322521	,6774793
3507	4,562	1,0000	,237383	,7626172
3514	4,135	1,0000	,308804	,6911955
3521	2,733	,6193	,162471	,4568294
3531	4,599	1,0000	,231079	,7689207
3541	2,829	,7018	,228922	,4728784
3570	4,742	1,0000	,207178	,7928218
3609	4,730	1,0000	,209260	,7907402
3622	2,002	,5283	,193648	,3346522
3631	2,750	,5122	,052488	,4597118
3634	2,077	,5556	,208405	,3471952
3636	4,740	1,0000	,207661	,7923386
3644	2,249	,5139	,137989	,3759105
3647	4,761	1,0000	,204056	,7959442
3656	4,875	1,0000	,184987	,8150131
3672	2,241	,5644	,189782	,3746180
3700	4,389	1,0000	,266206	,7337938

3701	2,288	,5294	,146869	,3825314
3708	4,000	,9286	,259961	,6686386
3712	4,831	1,0000	,192421	,8075788
a. Dependent Variable: Dropout				

Residuals Statistics^a					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	,001818	,366903	,189954	,0524542	3737
Std. Predicted Value	-3,587	3,373	,000	1,000	3737
Standard Error of Predicted Value	,003	,015	,004	,002	3737
Adjusted Predicted Value	,001831	,366875	,189934	,0524867	3737
Residual	-,3560863	,9316621	,0000000	,1671315	3737
Std. Residual	-2,130	5,573	,000	1,000	3737
Stud. Residual	-2,134	5,581	,000	1,000	3737
Deleted Residual	-,3573811	,9343112	,0000197	,1673092	3737
Stud. Deleted Residual	-2,135	5,604	,000	1,002	3737
Mahal. Distance	,000	28,074	1,999	3,085	3737
Cook's Distance	,000	,032	,000	,002	3737
Centered Leverage Value	,000	,008	,001	,001	3737
a. Dependent Variable: Dropout					

9. Reflection and philosophy of science

“Bringing a normal research problem to a conclusion is achieving the anticipated in a new way, [...]” (Kuhn, 1962, p. 36).

As a junior researcher I deem it important to reflect on the way I conduct my research, and philosophy of science – in the broad sense – is a suitable means to guide me through the process. In this section, I reflect on the current research as a whole, the thesis, and my position as a researcher, however this is not a position statement of my view (as a researcher) on the world and on what knowledge we can gather about it.

“The man who is striving to solve a problem defined by existing knowledge and technique is not, however, just looking around. He knows what he wants to achieve, and he designs his instruments and directs his thoughts accordingly” (Kuhn, 1962, p. 96). This quote made me laugh, both because of the common-sense logic that is behind it (even though these are Kuhn’s words) and the non-sense it is when looking at my own work. However, it started out that way – looking for a problem that needed to be solved, guided by the context of a larger research (The end of membership / Lid van de club), building on the existing knowledge of that larger research. But, as time passes it became clear that existing knowledge was almost negligible and that techniques were not all that straightforward either. At first, the idea was to do a – brief – statistical analysis and then to use those findings to conduct qualitative research, primarily interviews. However, as you will learn when you read the rest of this thesis, statistical analysis alone is more than enough – for now. And not only was the statistical analysis part more demanding than expected, there was a long road ahead just to get the right data, one of dialogue, as will be discussed below.

Miedema (2012) explains different forms of science in a more or less Kuhnian sense: first, there was science 1.0, which put emphasis on the researcher as being autonomous and research as science-driven. Then there was science 2.0, which entailed more dialogue with social stakeholders about the results and products of research. Even later, science 3.0 came about, geared towards co-creation in which scientists work in partnership with external parties in order to seek a solution for a problem. The current research is an example of science 2.0 as dialogue with social stakeholders is started cautiously. I say cautiously here, as the dialogue

could be more engaging and more leaning towards science 3.0, if there was only the time. However, something that caused me to hold back even more, that is caused me not to engage too much in a dialogue or co-creation, are the expectations the conventional curricula for master students have: conduct your own research, write up your own results. These expectations – along with some other developments in academia (Stapel) – caused me almost to be weary of (too much) discussion, arguments, and dialogue about my own research. In addition, I am very lucky to have had the opportunity to engage with my supervisors, my peers, and the KNVB in discussing relevant aspects of the data in perspective of my research interests. I think this is a shame, as dialogue and co-creation could help science to be more engaged in real world issues, and that real world issues bring about very interesting issues to research. Miedema (2012, p.9) states that “These more concrete matters – the problems of science and society – are easier to analyze and usually more interesting than more abstract questions involving intrinsic aspects of the scientific endeavor.” (I doubt the analysis of these problems is easier, but still, I think he has a point here).

Another interesting aspect about science 2.0 and science 3.0 is related to the famous ideas of Karl Popper, who writes in his *The Logic of Scientific Discovery* that verification does not exist, and that science should be attributed towards falsification. Conversely, Kuhn (1962, p. 147) makes an interesting point when he states that “[...] it is in that joint verification-falsification process that the probabilist’s comparison of theories plays a central role. [...] it may also enable us to begin explicating the role of agreement (or disagreement) between fact and theory in the verification process.” Even though Kuhn talks here about the natural or exact sciences, the idea behind it is very interesting for social scientists as well (or especially): as dialogue, and subsequent (dis)agreement can evolve into a better understanding of our research subject.

My academic education started with a bachelor in Cultural Anthropology, which thought me a great deal about interpretative and relational philosophical viewpoints, mostly based on interviews and participant observation. This particular field of studies is not very familiar with other methods such as surveys or databases for data gathering, or statistical methods for data analysis. The initial idea on my thesis research was to use statistical analyses to inform which cases should be researched on a qualitative basis. Unfortunately, the time frame for this thesis research did not allow for more in-depth qualitative inquiries of the research topic as datasets were harder to access than previously thought and statistical analyses claimed more time than reserved for them. As someone schooled in philosophies guided by relationality and

interpretation, this was an extra challenge: how to bring in the interpretation into something that is deemed rigorous and thorough, and how to understand relations between X and Y, knowing so much more could be included? Still, the research process helped me understand that even statistics is based on human decision-making, and that SPSS is just a tool. It is the researcher who has to make sense of all the output SPSS creates. And it is the researcher who decides what goes into SPSS in the first place. Reporting these decisions, and making decisions on the basis of sound argumentations is something I have learned by conducting this research.

