TV Tinder

Exploring how the use of a video match-maker system can potentially improve co-viewing experiences of online public television

Student Name: Jessica Broeders
Student Number: 432482

Supervisor: Dr. Elisabeth Timmermans

Master Media Studies - Media & Business
Erasmus School of History, Culture and Communication
Erasmus University Rotterdam

Master's Thesis
June 2020

TV TINDER: EXPLORING HOW THE USE OF A VIDEO MATCH-MAKER SYSTEM CAN POTENTIALLY IMPROVE CO-VIEWING EXPERIENCES OF ONLINE PUBLIC TELEVISION

ABSTRACT

This research investigated to what extent our mock-up version of a group recommender system – that was created for the purpose of this study – positively contributed to the co-viewing experience of two people (a dyad), regarding public service media. Through the theory on Social Uses of TV by Lull (1980) we have shed light on the oftentimes forgotten social context in the development of group recommender systems, and their underlying algorithms to predict accurate viewing suggestions. By building on this theory with empirical research in the field, we have arrived at our hypotheses. A quantitative cross-sectional survey research was conducted to gather our data from a sample that matches the target audience of the Dutch public broadcaster NPO. We have taken some inspiration from experimental research in designing our methodology to ensure that the presentation of our mock-up group recommender system (GRS) was as closely linked to reality as possible (in our case duplicating the existing interface of the NPO online streaming platform). We foresaw that a match or mismatch in the viewing interests of our dyads, and the indication of their social relationship quality would influence the intention to follow up on the recommendations produced by our GRS. However, as appeared from our results, these associations were not significant. In this study, perceived usefulness of the endorsed videos by the GRS was measured looking at three constructs that recur in the debate on creating recommender system algorithms, being accuracy, novelty, and diversity of recommendations. A Hayes parallel multiple mediation model was conducted on the mediating power of perceived usefulness on the relationship between the perceived relevance of the GRS (captured by further use intentions) and the watching intentions of the suggestions put forward by this GRS. The results from our mediation analysis tied in closely with the prevailing accuracy-diversity debate as it complicates the creation of a diversity-aware recommendation model serving the societal function of the public broadcaster. Only accuracy was found to fully mediating this relationship, however, we argue in our discussion that although the concept of diversity is harder to measure in terms of audience acceptance, it should not be underestimated for its valuable contribution to the richness of the media landscape in the Netherlands.

KEYWORDS: *Co-viewing, Group Recommender System, Public Broadcasting, Diversity, Video Endorsement*

Table of Contents

1. Introduction

Over the past decades, our daily lives have been increasingly filled with multiple options, editions, colours, updates, extras, and price categories to purchase in, as for a long time Western consumption society has been characterised by the same growth-based ideal. Ironically enough, consumers think they value the idea of having a choice, yet an overload of choice leads to symptoms of stress and being overwhelmed, in the worst case leaving them having made no choice at all (e.g. Bollen, Knijnenburg, Willemsen, & Graus, 2010; Gomez-Uribe & Hunt, 2015; Silveira, Zhang, Lin, Liu, & Ma, 2019). With more academic and practical awareness on the so-called choice paradox entering the field, the more the merrier idea is being slowly re-considered.

As an (indirect) result, recommendation systems have been in use for quite a while, attempting to make people's life easier in the overflow of options to choose from and decisions to make. Many systems narrow down extensive lists of options and learn to fit suggestions to your personality as they grow, based on richer data available; a process known as machine learning (Portugal, Alencar, & Cowan, 2018). Popular online streaming platforms such as Netflix and YouTube could be seen as the early adopters of personalized recommendation systems, continuously aiming to improve the accuracy of their recommendations. More recently, group recommender systems have been developed, after realizing many choice-processes are concerning more than one individual (Jameson, 2004). For instance, these systems may help a group decide which holiday to book, where to eat out, or what to watch while having a casual evening on the couch. These activities could be enjoyed unaccompanied, but are oftentimes an entanglement of multiple preferences, moods and opinions. In order to reach a decision that every group member stands behind, a process of negotiation and discussion may be necessary (Jameson, 2004; Lee, Heeter, & LaRose, 2010).

Similarly, television watching oftentimes becomes a social process where multiple people engage with the content provided to them on-screen (Tal-Or, 2019). Households increasingly have multiple television sets or screens for online streaming possibilities available, yet people still have a high tendency to enjoy watching television together – which is referred to as co-viewing behaviour (Mora, Ho, and Krider, 2011). Co-viewing, or group viewing, makes television watching interactive and engaging, facilitating them to create social and emotional connections (Chorianopoulos, 2007). This, for instance, takes place by discussing the content of TV shows with each other. With a shared viewing experience, however, comes the additional challenge of picking a television item to watch that everyone is pleased with. Surprisingly, many group recommendation systems (further: GRS) have failed to acknowledge that items recommended should also have the ability to appeal to a group instead of the individual. That is, certain individuals may be happy with the outcome, but that

does not mean that the group as a whole is satisfied. Especially in the case of repeated use, there is a change that 'the most average' group member(s) are better catered to than are the others, which can be a major cause for group dissatisfaction. Several researchers have pointed out that group decision making encompasses more than the sum of individual preferences (e.g. Jameson, 2004; Chaney et al., 2014), as will be elaborated on in the theoretical framework. Interestingly, previous literature on television viewing often excludes co-viewing, despite it being a more common practice than is individual television viewing (Mora, Ho, & Krider, 2011).

Relatively few studies have looked at the social aspects of co-viewing and their influence on the overall viewing experience (e.g. Tal-Or, 2019), let alone at the association between group endorsement systems and the viewing experience. Group endorsement systems are often misunderstood as a purely technological tool, forgetting the social process of the dual decision-making leading up to shared watching intention. Hence, the recommendations made and their accuracy (prediction rates) can benefit greatly from insights of societal processes (Mora, Ho, & Krider, 2011). This is where this study shows considerable scientific relevance, combining existing theories with new empirical findings. Some studies have focused on the mediated relationship between parent - child in co-viewing situations (e.g. Paavonen, Roine, Pennonen, & Lahikainen, 2009; Skouteris, & Kelly, 2006; Strouse, Troseth, O'Doherty, & Saylor, 2018), yet differentiation between various adult to adult relationships is missing (e.g. friendship, family, romantic). This study will hence dive a little deeper into the co-viewing experiences of people in a dyadic relationship, which we will further refer to as a dyad. Moreover, this study aims to tackle some of the most prominent limitations found in similar studies. The results of this research will provide insights into the reception of co-viewing endorsement systems, and the decision-making processes that are inseparably linked with it. Some of the challenges addressed by previous literature will be discussed in the context of the current study.

Since a large amount of activities that could be endorsed on an individual level, could also be performed by groups, the growing economic potential of group recommendation systems will be outlined by means of this study. This study provides insights in the acceptance of the GRS selection that has been shown to the participants, which could be duplicated to fit other business models. Furthermore, since the focus lies on the attitude towards the usefulness of the GRS, processes of mutual (dis)satisfaction could provide valuable insights on consumer co-viewing behaviour. Besides that, a better understanding of co-viewing patterns may also help to explain individual television watching behaviour. Likewise, the technology behind group recommender systems may be beneficial to solve issues related to single recommender systems that still make use of aggregations. For instance, this technology could contribute to what is known as the 'cold-start problem', when a system is just put into use and cannot base its recommendations on previous

user activity (Masthoff, 2004). According to Jameson (2004) and Bollen, et al. (2010) there is no guarantee yet that any of the recommendations by a (group) recommender system will be adopted, hence creating a model to assess the effectiveness of the selection of options will add to the practical implication of these systems.

On the other hand, according to Netflix, an estimated two thirds of their content is watched as a result of their recommendation system, providing several screens of personalized endorsements. This is saving them around 1 billion dollars per year due to increased user retention (Adomavicius, Bockstedt, Curley, & Zhang, 2013; Gomez-Uribe & Hunt, 2015). The potential to translate this business value to other sectors proves once again the value of investigating the next steps in recommender systems and their underlying algorithms. On a more practical level, this could provide video streaming platforms with a framework to assess whether the distance between preference and suggestion is regarded acceptable or not. For public broadcasters in specific, who oftentimes have an obligation to serve the public with a wide variety of content, this could indicate to what extent they are able to persuade their public towards watching a more diversified and 'out-of-the-box' set of programmes (Sørensen & Hutchinson, 2017). Hence, this research will provide an answer to the following research question: *To what extent does the use of a video recommender system, based on a dyad's combined viewing interest, positively contribute towards co-viewing experiences of online public television streaming?*

In order to do so, we will start with outlining our theoretical framework, which can be found in Chapter 2. The theory of Social Uses of TV by Lull (1980) – who in turn has based his theory on the theory of Uses and Gratifications by Blumler and Katz (1974) – will be used as starting point for developing our hypotheses, which will be further supplemented by empirical research. Chapter 3 will discuss the design choices for our methodology. By conducting a quantitative survey, several interfaces of an online public television streaming platform will be tested on the perceived usefulness of their personalized recommendations and the relevance of the mock-up group recommender system we have utilized. Subsequently, the intentions of users to watch their recommendations will be measured, along with the influence of social relationship quality on their watching intention. The co-viewing aspect of this study will be added by taking the viewing interests of others into consideration, creating combined genre preference with a matching set of recommendations. In the results chapter, the analyses and the outcomes of the proposed hypothesis are described. Finally, Chapter 5 will elaborate on these findings and address the practical and scientific implications that arise when linking them to relevant literature in the field. Furthermore, some of the strengths and limitations of this study, as well as suggestions for further research will be presented.

2. Theoretical framework

The most prominent theoretical discussions on the topics of recommender systems and co-viewing will be introduced. By combining some of this literature with new academic insights on group recommender systems, an outline of the current (social) challenges and opportunities in developing and assessing the effectiveness of these systems has been provided. Furthermore, we will touch upon the applicability towards public television broadcasting, leading us to present the five hypotheses that will further guide this research.

2.1 Recommender systems

In a Western consumeristic world where choices are ought to be unlimited, recommender systems are a new feature to facilitate better, quicker, and easier decision making. They are regarded key in retrieving information and/or discovering appealing content that is relevant for an individual in a digital environment where it is close to – if not completely – impossible to consume everything that is stored online (Davidson, Liebald, Liu, Nandy, & Van Vleet, 2010). Recommender systems act as a navigator in a sea of information and lead you to a satisfying piece of information or entertainment in an efficient way. Albeit a common argument, the authenticity of this objective for creating recommender systems has been challenged in the past. According to Sørensen and Hutchinson (2017), the main reason that recommender systems have been introduced is to ensure optimization of exposure to programming of the medium's choice and are not necessarily in line with favourable audience preferences. Despite the conflict of interest, the benefits of such systems are generally well-established. For instance, better focused audience attention as a result of using a recommender system would often translate into users paying closer attention to the content shown, and longer watching time, which contributes to the medium's exposure (Mora, Ho, & Krider, 2011).

One of the first implementers of recommender systems is the field of e-commerce, in an effort to uplift online shopping experiences by using predictive mechanisms (Adomavicius, Bockstedt, Curley, & Zhang, 2013). From a business perspective, this has considerable advantages concerning customer service and -need fulfilment. Being able to build upon realistic prognosis of what your client wants, enables a business to tailor its service towards the ideal customer journey, ultimately increasing sales. It is said that recommender systems can be responsible for about 10% to 30% of the total sales of online retailers (Adomavicius et al., 2013).

A recommender system can be defined as "a tool used by websites to facilitate its users in locating targeted artefacts, such as, articles, music, movies or products" (Xue, Moitra, & Gustafson, 2013, p.6). More specific for our research, Davidson et al. (2010) describe *video* recommenders as

systems that endorse a personalized selection to help users along the way to what they would (ultimately) perceive as high-quality content. In order to achieve this, it is important to keep users engaged and up to date, for it can benefit both company and consumer to portray a wide variety of the platform's availability in the long run. In addition, video recommenders are designed to predict which unseen video is ought to receive the highest user ratings, and endorse those (Xue et al., 2013). In most cases, these endorsements are personalized so that not everyone will see the same set of videos selected by the recommender system. Generally, the input for recommender systems consists of user ratings of previously watched items as a tool to find recommendations for future sessions. These can then be improved using techniques such as data mining or machine learning. By estimating preferences based on the user's liking of items yet to be consumed, the user ratings are turned into system ratings. Following up, accuracy is then measured and improved by means of a feedback loop consisting of the actual user ratings of estimated preferences (Adomavicius et al., 2013). Again, these circular consumer ratings are the most common way to evaluate the recommender system's performance. Yet, just as each recommender system is based on different inputs, needs, and/or user activity, finding an effective way to measure its success may differ depending on the setting or its expected user-supporting task, which may in turn influence the embedment of the user experience format and subsequent choice of algorithm (Ekstrand, Riedl, & Konstan, 2011).

A helpful technique to reach these estimated preferences contains *collaborative filtering*, which is a way of recommending items based on what rating other (similar) users gave to those items (Ekstrand, Riedl, & Konstan, 2011). Additional user-based activities that are able to compliment this technique include the percentage of finished shows, sequences watched, related content, but also diversity approaches can be implemented (Davidson et al., 2010). This is considerably different from *content-based filtering* approaches for recommendation, which concerns the user's own activity and liking of content, and finds similar content by solely looking at the metadata (textual cues) attached to those items (Ekstrand, Riedl, & Konstan, 2011). The benefits of using a content-based approach using metadata in relation to public broadcasting recommending will be discussed later (see: section 2.3.2 of this Chapter).

Whereas many studies have focussed on improving current recommendation systems' accuracy and predictive features, Adomavicius et al. (2013) argue that reliability of those systems may be of more importance than their precision. Concluding the experiments by Adomavicius et al. (2013), the recommendations portrayed by the system functioned as an anchor for the viewing preferences of the consumers. The use of recommender systems has an underlying behavioural aspect that is often not considered when building the algorithms that comprise them. Typically, an algorithm is comprised of a set of rules that act as gatekeepers for selection and recommendation

of certain media content (Sørensen, & Hutchinson, 2017). The complexity of those rule-based systems is often increased by its non-transparent nature – it is generally not known what the algorithm is constituted of, partly accelerated by machine learning (Portugal, Alencar, & Cowan, 2018; Sørensen & Hutchinson, 2017).

According to behavioural decision theory, it is common for people to be influenced by and have preferences formed by their environments – in this case the persuasive element of recommender systems (e.g. Adomavicius et al., 2013; Köcher, Jugovac, Jannach, & Holzmüller, 2019). That is, anchoring effects occur, which means that an individual is drawn towards recommendations that are shown to them at the time of consumption (Zhang, 2011). According to anchoring effects theory, a person considers an initial value (such as the recommendations by the recommender system), which he or she then processes internally and adjusts as wished to arrive at a final conclusion (Adomavicius et al., 2013). This conclusion is often skewed towards the initial anchor (value), and therewith comprises a systematic bias in common judgements. In accordance with these effects, Adomavicius et al. (2013) have found that recommendations by a recommender system can strongly impact the preferences of its users, indicating that one's preference ratings are malleable and are continuously drifting closer towards those of the recommendations on screen. This indicates that general recommendation systems may be able to construct the preferences of their users.

Furthermore, many recommender systems are built to enforce pre-existing views and preferred content, which makes it much easier for users to accept the content proposed by the system, as they tend to agree with our own (this is generally known as the accuracy of recommendations; Tintarev, 2017). However, recent debate fostered accusations of recommender systems leaving out content that is needed to assist in broadening people's horizon. These accusations are generally gathered in the concepts of echo chambers or filter bubbles. Filter bubbles refer to the consequences of filtering data streams so that recommendations are matched and personalized to one' interests and/or perspectives, complementing people's preference to see content that they agree with or that appears familiar (Nagulendra & Vassileva, 2014). On the one hand, these filter streams enable people to find their way in an overload of information. On the other hand, it carries the risk of filtering out important social or cultural functions of media, facilitating the creation of echo-chambers for repeating one's self-constructed reality. Ultimately, this could make people dangerously unaware of critical and/or minority voices (Nagulendra & Vassileva, 2014). The act of filtering content is not just one that can be solely attributed to recommender systems; users too, articulate different diversity needs in their personal lives and not all content will satisfy individuals' needs similarly (Tintarev, 2017). Thus, in order to maximize the fulfilment of their needs, personal filtering takes place as well. This study will provide more insights

into the interplay between the wish for accuracy in recommendations, and to what extent recommender systems can help in bursting the filter bubble by providing diversified content.

2.1.1 Video endorsement through recommender systems

In order to identify how videos are endorsed through popular online recommender systems, Davidson et al. (2010) examined the algorithm behind YouTube recommendations. Generally, the two most frequent motivations to use the platform are a goal-oriented search for a video or topic, or a more cognitively originated search for entertainment. The latter is referred to as an *unarticulated want* (Davidson et al., 2010). The latter is where recommender systems can contribute the most, yet they may need the input of the former to make accurate suggestions. YouTube, in specific, uses a personalized approach to video endorsement based on one's activity on the platform, together with querying (entering a specific search term) and (non-)directed browsing through the 'Browse' page. In more general understanding, YouTube video recommendations are built upon three pillars (Davidson et al., 2010). The first one is video quality, which describes the appreciation rate of the content across the platform, despite the interests of the individual user. Secondly, user specificity is linked to a user's personal preferences and past viewing behaviour. Here, we can see that YouTube employed the collaborative filtering approach within their recommender system. Thirdly, the platform makes use of diversification, which ensures a small list of relevant yet diverse set of recommendations, as space for showcasing them is limited. Videos that are too similar, and hence would generate a very narrow line of recommendations, are replaced with actual new content to the user (Davidson et al., 2010). As we will see later in this work, diversification of recommender systems has been central to many discussions. Hence, we can learn from the diversification approach YouTube has created and which it has balanced well with the relevance of the recommendations to the users.

However, diversity can take up various forms (other than replacing too similar content), as well as its execution may vary across the intended goals of the platforms. That is, in many cases the recommender system is ought to fulfil more than one criterion. Some examples would be the relevance of the recommendations based on users' recent activity (accuracy), whether the endorsed items are something that the user had not seen before (novelty), and a logical understanding of the rationale behind given suggestions (transparency) (Tintarev, 2017; Zhang, 2013). Depending on the platform and/or their commercial interests, diversity may of bigger or lesser importance.

To illustrate, one recommender system that has articulated the goal of its recommender system substantially well is that of Netflix, as became apparent in the launch of 'The Netflix Prize' late 2006 (Bennett & Lanning, 2007). The Netflix Prize was a challenge addressing anybody who

thought they could beat the recommender system that was employed at that time, Cinematch. The anonymized dataset used for this, is now openly available for the purpose of non-commercial research. The Netflix recommender system is famous in its field for what has been named 'the Netflix experience', personalized content for multiple settings, provided to the user by the combined work of various algorithms (Gomez-Uribe & Hunt, 2015). The complicated and intertwined nature of this technology will be well-suited to build group recommender systems that take into account preferences and user activity of multiple users simultaneously, as will be addressed later in this paper (see section 2.3). Netflix is employing the latest machine learning developments in algorithm development technology, taking collaborative filtering as a basis, yet there is still room for major improvement (achieved recommendation accuracy of 0.85, see Adomavicius et al., 2013).

It should be noted that at the time of the competition, Netflix was still a DVD rental service, based on a monthly subscription fee. The goal of the Netflix recommender system, however, can still be translated to its present-day functioning: to spark the interest of subscribers fast enough to find those movies and/or video content that they will certainly enjoy. Gomez-Uribe and Hunt (2015) pointed out that the attention span of users is at the core of the well-workings of a proper recommender system. Specifically focusing on the Netflix platform, it was investigated that viewers generally spent one minute to a minute and a half searching for a programme to watch, before the chance of abandoning the platform drastically increases as a result of lost interest (Bennett & Lanning, 2007; Gomez-Uribe & Hunt, 2015). According to Gomez-Uribe and Hunt (2015), it is of great importance to design a recommender system in such a way that within this limited time span, the user is able to review a compelling set of recommendations and understands why these are relevant to him/her specifically (transparency criterion). In public service broadcasting, as a result of the obligation to provide the public with a diversified portfolio, providing a rationale behind recommending items is regarded even more important (Sørensen & Hutchinson, 2017). Besides that, public broadcasting still has to be responsive to the preferences of the public. As Sørensen and Hutchinson (2017) describe, as so-called 'gatekeepers' for disseminating knowledge, not showing transparency about the logic of recommendations can foster accusations of paternalism – which in this case entails the interference with audience preferences, and concerns distribution of programming for the public broadcaster's own good. This is referred to as the paternalism – popularity debate, which is also highly applicable to the Dutch media landscape (Sørensen & Hutchinson, 2017; Bardoel, 2003). Here, the interplay between the different criteria of transparency, accuracy and diversity can already be identified, which once again shows the challenges that contemporary recommender systems face. This study will proceed with identifying and discussing the implications of co-viewing and group recommender systems respectively, for

which these issues still apply or become even more apparent.

2.2 Co-viewing

In academic discussion on recommender systems – or even group recommender systems – co-viewing is often not taken into consideration, although it being one of the most frequent and most natural behaviours when it comes to television watching (Mora, Ho, & Krider, 2011; Tal-Or, 2019). For that reason, we believe investigating the one goes hand in hand with addressing the other, thus we will provide substantial attention to the topic of co-viewing and the implications it has on the shared television watching experience. In the next section (2.3) we will dive deeper into the consequences of co-viewing for developing group recommender systems.

Chorianopoulos (2007) has written about the social uses of television watching in a digitalized and interactive TV context. Here, one can distinguish two categories: connecting physically distanced people through the use of interactive TV (both synchronous such as communication through digital platforms at the time of watching, or asynchronous such as engaging in discussion about TV content all members have seen previously) and co-viewing. The terminology co-viewing is used when referring to the shared experience of joint television watching from the same screen, generally by small groups of affiliated people (e.g. friends and/or family; Chorianopoulos, 2007). It should be noted that this is significantly different from 'social TV', which aligns audio-visual systems with an interactive feature enabling distant users to communicate interpersonally (Chorianopoulos, 2007; Bellman, Robinson, Wooley, & Varan, 2017). An example of the latter would be the Google Chrome extension called Netflix party. This extension allows Netflix users to bridge their physical distance by synchronizing the videos they are watching, including pausing and replaying (Young, 2020). Co-viewing, on the other hand, does not need any technological guidance per se, as it is concerning viewers who are in the same room.

Considerate academic research has looked into co-viewing from the lens of the family set-up. Television watching within a family was found to be merely convenient behaviour that is most easily achieved together (Mora, Ho, & Krider, 2011). Despite the increased challenge to find a programme that fits all viewing interests, it is considered a sociable and cheerful activity. Engaging in shared television watching behaviour would also be highly beneficial for the family bond, as co-viewing has the ability to resolve conflict and increases solidarity in a family (Mora, Ho, & Krider, 2011). The remainder of this study will be focusing on co-viewing practices, and thus physical togetherness, of all kinds of adult to adult relationships. Studies on co-viewing where parent-child relationships are at the centre are common (e.g. Paavonen, Roine, Pennonen, & Lahikainen, 2009; Skouteris, & Kelly, 2006; Strouse, Troseth, O'Doherty, & Saylor, 2018), but as they require a completely different angle emerging from a child psychology and parental influence standpoint, we

will not further examine this. Before continuing to discuss the social uses and implications of co-viewing, it should be pointed out that many of the existing literature on this field is written from a highly westernized (in many cases Americanised) perspective. Thus, it would not always be appropriate to apply the literature and/or findings outlined in this work towards a global scale. Since we are aiming to continue the discussion in a similarly westernized context, we do not foresee any major complications.

2.2.1 Social uses of TV

Recent debate stirred up the discussion around technological developments in the field of social engagement with television watching. When it was first introduced, the television was put at the centre of technologies that harm social interactions among people. This because it would divert attention away from authentic face-to-face communications as it was known previously, both applicable to the family setting, and local community interactions (Chorianopoulos, 2007). The idea was countered by behavioural scientists who looked into the television as a social medium, with benefits for bonding with acquaintances or even strangers for the 'common points of reference' it creates (e.g. Girgensohn & Lee, 2002; Lee & Lee, 1995). Instead of a decline in civic engagement, Chorianopoulos (2007) argues that television viewing paves way for group viewing, ultimately leading to shared emotional responses and experiences, and the process of building common ground amongst a group. This is driven by the basic social instinct of human beings to interact and enjoy discussions about shared interests. Although it has been well-established that co-viewing is a common social activity for shared households, disagreement in the academic world exists about the future of joint television watching. For instance, Chorianopoulos (2007) puts current societal developments central to the increased difficulty with which he claims the intention to engage in co-viewing is burdened, including the stressful events of daily life and growing scattering of people among households. On the contrary, Mora, Ho, and Krider (2011) argue that an increase in single-person households and the availability of multiple TV-accessible screens is on no level affecting the amount of co-viewing adults – which was in fact growing in the UK.

This desirable social aspect of television watching could be explained by the theory of *Uses and Gratifications* by Blumler and Katz (1974), which explains media consumption through a need for social interaction with others (Haridakis & Hanson, 2009). This theory lies at the basis of the *Social Uses of TV* by Lull (1980), which we touch upon later, hence we will discuss the theory of Uses and Gratifications shortly. The motivation to watch together is a natural tendency to satisfy several needs, such as intimacy, fostering communication, and justification of one's competency (Tal-Or, 2019). Rubin (1983) describes several typical uses and gratifications specified towards television watching, which would again have the power to influence an individual's attitude and

perception of the television content. Among these uses and gratifications are the building of one's personal identity and the fostering of relationships between viewers. Mora, Ho, and Krider (2011) argue that some of these social needs can only be fulfilled through the process of co-viewing, such as building a shared agenda for conversation. Additionally, the companionship need is processed as a higher emotional reward as would be the informative or entertainment aspect of TV viewing. Interestingly, media is used for satisfaction of interpersonal needs, yet interpersonal communication can also be beneficial to satisfy media intake (Haridakis & Hanson, 2009).

2.2.2 Individual vs. group television watching

Watching television alone is considerably different than watching television accompanied by others (Haridakis & Hanson, 2009). As Tal-Or (2019) points out, human beings possess limited capabilities to focus our attention. With regards to co-viewing, this implicates that as television watching became a social activity, there are more factors to divide our attention towards. This would hinder the process of transportation (being fully absorbed into the narrative, away from a reality context), which again could affect overall engagement with the show (Tal-Or, 2019; Lee, Heeter, & LaRose, 2010). Similar results were found for understanding of difficult parts of the show, such as certain dialogue. It is plausible that at the moment of deciding on a programme to watch together, being aware of this existing attention divide has an influence on the individual's viewing preferences. A show that is hard to follow or would require keeping up with the dialogue would logically have lower preference in a co-viewing situation. This could result in a different co-viewing experience than merely summing up preferences based on both individuals' watching history. In addition, Lee, Heeter, and LaRose (2010) pointed out that reduced story involvement could be a tactic to keep group members satisfied about their restricted ability to choose, keeping the cohesion of the group intact. This may indicate that people care less about the outcome when there are more factors involved in the choice process, in order to keep the group enjoyment of the show high (Lee, Heeter, & LaRose, 2010).

On the other hand, as Zajonc (1965) indicated through his idea on social, co-viewing can in some cases intensify the arousal and emotions with which is watched. This was established for children's TV shows but may also hold true for adult to adult relationships (Tal-Or, 2019; Skouteris, & Kelly, 2006; Strouse, Troseth, O'Doherty, & Saylor, 2018). Moreover, watching certain programmes together can create an image of fondness towards the other and his/her choice for the content shown, which will positively reflect on the attitude of the other person towards this type of content (Tal-Or, 2019). This could indicate that the distance between interpersonal viewing preferences could be overcome due to psychological processes of appreciation for the other and his or her choices. Likewise, people may have an interest in knowing what the preferences of the other

person are to base their own decision on (Jameson, 2004). This could result from two principles, either saving of effort or learning from each other. The principle of learning from each other is based on the idea that people in an undefined relationship are interested in discovering what the other person is passionate about, if they did not know beforehand (Jameson, 2004). In sum, people are naturally willing to engage in learning processes about the other, which may reflect on a liberalized approach towards the final co-viewing decision. In the end, human beings are inclined to minimize conflicts between people (Jameson, 2004).

By linking both the theory on uses and gratifications approach by Blumler and Katz (1974) and findings on the topic of co-viewing, Lull first introduced his theory on the *Social Uses of TV* in 1980. Lull (1980) regards television watching, and more specific co-viewing, as a source for real-time communication between viewing partners, and above all a conversation-starter (Geerts & De Grooff, 2009). Lull (1980) discovered that co-viewing was dependent on the psychographic profiles of viewers (their personal traits, habits, values etcetera), and that higher similarity leads to higher co-viewing intentions. As psychological characteristics influence one's preferences, this implies that diverse couple profiles would have different taste in television programmes and/or genres. Consequently, this couple would be less intended to engage in co-viewing behaviour, or would require more conflict resolution prior to co-viewing in order to reach a consensus on what programme to watch together (Lull, 1980 as cited in Mora, Ho, and Krider, 2011).

In short, co-viewing is said to arouse a wide variety of effects that are not as strongly showing when watching alone, such as higher levels of engagement with a show and higher viewer attention, which can both be explained as a result of interpersonal conversating elicited by the act of co-viewing (Mora, Ho, & Krider, 2011). Interestingly, Mora, Ho, and Krider (2011) add to this that the anticipating thought of engaging in co-viewing behaviour is likely to increase the enjoyment of a series one is about to watch. Furthermore, longer watching times and less browsing through channels to find the right pick are consequences of so-called inheritance effects, which indicate that it would ease the group choice process to stay on the same channel for the remainder of the evening. This once again shows that people in a group tend to minimize conflict, which may yield some interesting outcomes when applying this to the creation of group recommender systems.

2.3 Group Recommender Systems (GRS)

Now that we have outlined some of the most recent debate on the topic of co-viewing, it is even more surprising to see that group viewing behaviours have been continuously left out of the prediction system development (Mora, Ho, & Krider, 2011). Following the findings by Haridakis and

Hanson (2009) that co-viewing can predict exposure to online videos just as much as it can for traditional broadcasting, it would make sense to apply co- viewing behaviours to group recommender system creation in an online streaming environment. Likewise, Chorianopoulos (2007) concluded that (group) video recommendations generally apply to online video streaming. Hence, we will address and dive further into the discussion on GRS from an online and on demand perspective, where applicable focused towards television broadcasting in specific. In this context, on demand refers to the ability to pick and watch various television programmes of interest wherever and whenever one pleases, possibly for a registration fee (Tryon, 2015).

2.3.1 Group recommender systems and group aggregation

In addition to the definition of individual recommender systems, group recommender systems aim to reach a joyful experience of the recommended items by all members of a certain group combined (Ekstrand, Riedl, & Konstan, 2011). Derived from the challenges to build algorithms according to different personalities and insert those into personal (individual) recommender systems, this brings along similar complications in building recommender systems based on group interactions. There are many limitations and challenges linked to group recommendation systems in their current existence. Jameson (2004) and Chaney et al. (2014) argued that there is a major difference in the functionality of the systems, compared to single recommender systems, since a GRS is not built on the simple sum of individual viewing preferences. There are usually two ways in which aggregation (combining) of preferences can be adopted; individual preferences are combined into group preferences on which the suggestions are based, or individual item suggestions are combined into group suggestions (Masthoff, 2011). However, only a limited amount of systems has been able to move towards an aggregation function with little room for users to manipulate the group recommendation – for instance by choosing more extreme viewing preferences to shift the average to their advantage (Jameson, 2004; Masthoff, 2011). Masthoff (2004) investigated how group voting behaviour could be interpreted and converted towards a group recommender format, based on studies originating from collective decision-making. She outlined several complications by means of several possible strategies, such as the chance of leaving some people highly miserable or treated unfairly compared to others, which were also the highest concerns from a user perspective (Masthoff, 2011). The decision on which of those strategies to employ should be done carefully, taking into account the context of the group decisions and the possible consequences of misery (e.g. higher chances of abandoning your platform). Possible solutions to misery would be strategies on rotation, where the person with less influence in the choice process would gain more saying power in the next round of usage of the group recommender system (Masthoff, 2004). According to Masthoff (2004), as the complex group interactions only apply to a much lesser extent to small

groups or dyads; for dyads the most common strategy is the Average strategy. Still, in its current form, researchers should be careful in attempting to duplicate group voting strategies to television viewing or co-viewing settings, since streaming platforms generally broadcast too many shows for people to list and rank all of them. As will be elaborated on in the next chapter (see section 3.2), due to the restrictions of our simplified mock-up version of a GRS, this limitation did not apply to our research.

Similarly to what Jameson (2004) and Chaney et al. (2014) stated, complications could arise too if current algorithmic filtering techniques are simply combined into an adjustment for groups. Yet, since specific technical issues of group aggregations and algorithms structures are beyond the scope of the design of this study, these will not be included in our theoretical framework. However, certain consequences of inserting algorithms into GRS's will help to elucidate the challenges public broadcasting companies are facing when considering the introduction of a recommender system. Similarly, we should also take a look at the social relationship between dyads in a co-viewing setting, since this may be applicable to the subjects of our study too. Thus, the following paragraphs will dive deeper into these elements, deliberately leaving out the technical complications. It is important to mention this, as it is at the base of constructing a proper GRS and brings along its own limitations and challenges that cannot be seen separate of the development and assessment of a GRS. Yet, since our focus is surrounding the assessment of a GRS in a public broadcasting and diversity context, these technological concerns would be applicable to a later phase of research.

2.3.2 Decision-making algorithms in the public service arena

Where algorithms are said to shape how information is circulating among people, they have an increasingly important role in public broadcasting given their agenda-setting and educational function in society (Sørensen, & Hutchinson, 2017). Sørensen and Hutchinson (2017) outline some challenges in embracing recommender systems, keeping in mind the core principles of public service media, of which the key questions raised are summarized below.

The balance between increasing the reach of public service media programming – hence serving the audience with content they liked and engaged with previously – and providing distinct and diverse content is one that has evoked discussion in the past (Sørensen, & Hutchinson, 2017). Especially with the rise of recommender systems the urge to choose either side grows, as the system is fed by regulations that are entered in the form of algorithms, which require a certain level of standardization for the system to work with. In contrast, diversification and uniqueness instead of standardization of content, is ought to be the main goal of a public broadcaster, giving rise to the existing debate (Sørensen, & Hutchinson, 2017). Interestingly, Tintarev (2017) argues that recommender systems carry both the ability to diversify and narrow down the content

recommended. Yet, as pointed out by Sørensen and Hutchinson (2017), increased complications arose with the entering of intermediate social network platforms (e.g. Facebook) who control most of people's attention span and interest. Thus, for public broadcasting companies, it becomes increasingly hard to compete with both accuracy-based and time-efficient recommender systems.

As mentioned earlier, both algorithms and human beings carry filtering biases that can influence the well-workings of any recommender system. An important premise is that when recommendations are combined in a model that is too simplistic (e.g. resulting in recommendations far off the user's preference), chances are that the user's intention for further use of the systems drop. Tintarev (2017) points out that it can be beneficial to study user perceptions as they can help to identify how presentational strategies can be shaped to foster acceptance of diverse content, as these strategies are said to be capable of mitigating the effects of challenging recommendations. The interplay between diverse and novel items and recommendations of interest is illustrated here. Again, the importance of transparency on recommendations proves fruitful, as showing the user his or her blind-spots (unfamiliar or underexplored areas of content) can lead to a higher encouragement of exploring these items, without damaging the system's reputation. The perception of discovery of these items increases its attractiveness for the user. However, there is fine line between diversified content that is perceived as newly discovered and diversified content that is not; the latter being a lot less attractive to the consumer (Tintarev, 2017).

Having touched upon the biggest differences in commercial and public broadcasting, and their relation to discussions on algorithm creation, we will otherwise presume that findings in the field of general television recommendations are applicable to public service broadcasting too, in absence of previous literature on our topic of interest. The following hypotheses will be based on academic theory and empirical findings from a broader GRS expertise, which will be applied to a public broadcasting context, to see if they still hold true in that arena.

2.3.3 Social relationship quality

Oftentimes, in the process of building a GRS, the social context seems forgotten, yet it is an important contributor to the acceptance of the proposed model of suggestions (Tal-Or, 2019). Social mimicry, conformity, and social cognitive theory predict that one's viewing experience is similar to that of a co-viewer (Tal-Or, 2019). This is an important reference point throughout the rest of this study, but this does not indicate that all groups are alike. Masthoff (2011) made a distinction between active and passive groups, and their level of interaction with the recommender system, which again affected the decision-making process in the end. Generally, the (G)RS assumes people are merely passively oriented and thus have limited interaction with the system. Moreover, in a study by Hennig-Thurau, Marchand, and Marx (2012), GRS outperformed single recommender

systems in terms of its derived value, and this effect appeared to be stronger when the social relationship was valued higher. The more liked and attractive the co-viewer, the more effectively he or she can draw the other in and out of the narrative (Tal-Or, 2019). In this light, the relative power a close co-viewer has over the other could transmit to a higher willingness to engage with the content provided. This indicates the importance of social relationship quality within groups; hence, we assume that the relational level will be contributing to the relevance of this research. We could argue that if the social relationship quality is valued higher, the dyads (group of two adults) are more likely to watch the recommendations outlined by the GRS. This is in line with what Jameson (2004) and Tal-Or (2019) described as learning about the other and being more highly appreciative about distinct programme choices as a result of the close relationship between people.

Still, the GRS may be more relevant for those with different viewing interests, as we expect that similar viewing interests will yield more easily acceptable viewing suggestions anyway, since negotiation will not be necessary. In the end, a consensus among viewers has to be reached and the bigger the difference in preferences, the more difficult the task of conflict management (Mora, Ho, & Krider, 2011). According to our main theoretical starting point, the social uses of TV, we described how more similar psychographic viewer profiles foster co-viewing intentions (Lull, 1980). People with similar psychographic profiles may share their outlook on many things, but could still portray very different viewing preferences. Here you can notice a slight differentiation between shared viewing interests and the nature of the relationship between a dyad. By means of the following two hypotheses we will test whether a match or mismatch in genre preferences and social relationship quality have an influence on the watching intention of the suggested content:

> **H1:** Compared to those dyads who entered different viewing interests, dyads who entered similar viewing interests in the GRS will indicate a higher watching intention of the selection of recommended videos.

> **H2**: Higher social relationship quality will positively influence watching intention of the selection of recommended videos.

2.3.4 Group recommender systems and effect studies

In absence of a theoretical model surrounding the effects of co-viewing, many researchers have been focusing on assessing co-viewing experiences *after* exposure to the full storyline (e.g. Tal-Or, 2019). Yet, according to the theory on *planned behaviour*, behavioural intention can be an excellent predictor of actual behaviour, making measuring the effectiveness of group recommender systems by means of watching *intention* a proper evaluative tool (Pu, Chen, & Hu, 2011; Venkatesh, Morris,

Davis, & Davis, 2003). Oftentimes, the evaluation models for individual endorsement systems are replicated to those targeting groups (Trattner, Said, Boratto, & Felfernig, 2018). However, as the social aspect of co-viewing indicated, the awareness of having to make a mutual decision might influence viewing preferences and watching intention of a group. Quijano-Sánchez, Díaz-Agudo, and Recio-García (2014) initiated a social recommender system that is based around knowledge management, which is said to improve as the system learns more about user satisfaction in the long run. Improved knowledge management would be a tool to increase the usefulness of the system, and hence paves the way to higher satisfaction rates (Quijano-Sánchez, Díaz-Agudo, & Recio-García, 2014). With regards to public service media, Silveira, Zhang, Lin, Liu, and Ma (2019) and Tintarev (2017) consider accuracy, novelty and diversity of the recommended items as the three pillars constructing the usefulness of a recommender system. Satisfaction with the recommendations provided (and thus, a system that is perceived useful) would then likely influence the intention to watch the endorsements. Accordingly, the following hypothesis will be tested:

> **H3:** Higher perceived usefulness of the GRS' selection of recommended videos will increase users' watching intention.

At this point, we can argue that the implementation of GRS could be improved, according to some of the limitations outlined above, but that the social value of this type of recommender system has been established. Even though, most group recommender systems simply activate the best match found and are not actively asking for agreement from the users, they generally perceive it as a welcoming tool (Jameson, 2004). In a field research by Jameson (2004), 80% of the people in the group requested the use of a GRS in a co-viewing situation, in order to allow the one person in charge of the TV remote to make a better final decision for them. Hence, it is assumed that;

> **H4:** Higher perceived relevance of a GRS (use intentions) will positively influence the perception of usefulness of the GRS' selection of recommended videos.

Lastly, assuming significant statistical proof in favour of hypotheses three and four will be found, a fifth hypothesis will emerge, combining the two statements into one. This hypothesis is in line with the limitation as pointed out by Horenberg (2019), who considers usefulness of the recommender's selection as a mediating factor in the analysis between the use of a GRS and the ultimate decision that is made (here: which endorsement to watch). Here, it would be expected that the use of a GRS, and further intentions to use a GRS would yield similar results.

H5: Higher perceived relevance of a GRS will increase the watching intention of the selection of recommended videos.

2.4 Conceptual framework

The hypotheses as outlined above can be logically interpreted according to the conceptual framework in Figure 1. All relationships in this framework are positive; meaning that once the value of the independent variable increases, likely will the dependent variable. One should take in mind that different results are expected among the experimental groups and the control groups – which will be elaborated on in the next chapter.

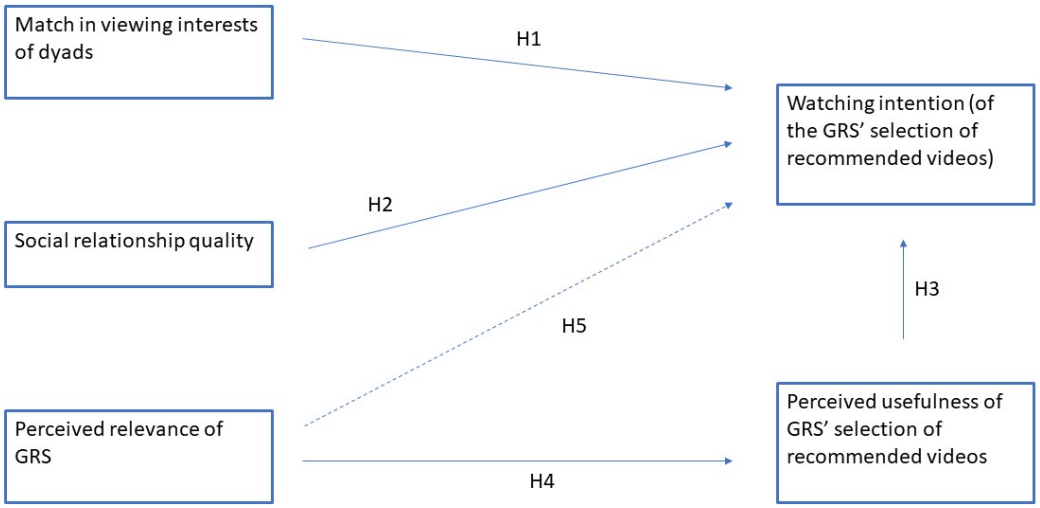


Figure 1. *Conceptual framework*

3. Methodology

This chapter will elaborate on the methodology used to answer our research question, and hypotheses. The choice for a quantitative cross-sectional survey design will be justified, followed by an extensive description of the procedure and the rationale in overcoming certain complications in the design. The technique for data collection will be touched upon shortly, after which the sample will be reported. Lastly, the measurements constructing the hypotheses will be outlined, including factor analysis and tests for reliability if needed.

3.1 Research design

In order to answer the research question at stake (*To what extent does the use of a video recommender system, based on a dyad's combined viewing interest, positively contribute towards co-viewing experiences of endorsed online public television streaming?)*, quantitative cross-sectional survey research was conducted. Matthews and Ross (2010) mentioned that quantitative methods are suited for those type of research questions where the researcher is already familiar with what they are looking for, which is expressed by means of our hypotheses. In fact, we are investigating a positive relationship – the influence of a group recommender system on co-viewing experiences/watching intention of endorsed videos – which is generally approached quantitatively (Neuman, 2014). Shelby (2011) explained how different segments or groups of people are likely to have varying answers in social science research, which made survey research a more suitable method for analysis since they require data from a bigger (representative) sample, that is more likely to include a multitude of perspectives. In turn, a bigger sample oftentimes accounts for better generalizability of the outcomes to the population, especially in the case of structured surveys based on closed-ended questions (Matthews and Ross, 2010).

With regards to group recommender systems, Masthoff (2011) suggested that experiments are the best way to assess whether group aggregations strategies work, because it provides the researcher with the ability to form several groups. Creation of groups may not have been as beneficial to our survey research, but it did give our participants the impression that they can choose, just like they would when engaging with an actual group recommender system (GRS). Even though the method we employed is not an experimental design, the advice from Masthoff (2011) was welcomed by creating a survey flow which allowed people to assign themselves to groups, improving the authenticity of the mock-up version of a GRS we have employed. Painting a picture that is closer to reality helps to better represent actual societal processes, which will likely have practical benefits for the generalizability of the study (Neuman, 2014). Although the respondents were directed in different directions, and shown varying screens as input, the design of a traditional

survey was followed by asking each participant the same questions about the varying interfaces in the same order.

3.2 Procedure

The research looked into the co-viewing behaviour of dyads and decision making with regard to intentions to follow up on/watch endorsed videos. Complete randomization of subjects by putting two people together without any (knowledge on the) relationship between the two could raise major concerns for the practical implications of this study. That is, group recommender systems are generally a tool invented to facilitate decision making amongst a social group with relational ties – people that are not complete strangers to each other (Hennig-Thurau, Marchand & Marx, 2012). Even though most research only looked into married partners, other forms of relationships between dyads may yield interesting results too. The mock-up version of the recommender system has been created to assess its usefulness by representative dyads and their intention to watch the recommendations that the GRS puts forward. Since this process is a shared effort, it would be beneficial to conduct the survey research in duality as well. Hence, based on the selection from a few exemplary programme genres, the dyads will 'assign' themselves to a certain group. Nor the dyads, nor the researcher have prior knowledge on the group they will end up in, so this process will be random. Both individuals in the dyad will pick a genre that has their personal preference at the moment of participation, whereupon the survey technology will redirect them to the right part of the questionnaire. The so-called flows that the survey can follow will be pre-entered based on possible choice combinations between genre A: Entertainment, B: Drama series, and C: News & current events. This leads to the following combinations resulting from two people's genre preferences: AA, AB, AC, BB, BC, and CC (6 conditions). Each combination will unlock a different screen where three programme suggestions will be shown, including a short description. Thus, in total there will be six different screens, of which one will be shown to the respondent, based on what genre he/she and the other person in the dyad picked before. The six screens can be found in appendix A. This approach has been chosen due to complexity issues, and the inability to create an algorithm for the sake of this research. As previously pointed out by Masthoff (2004), the average strategy on decision making is most common among dyads, hence combining genre preferences of two individuals aligned well with this approach.

Partly due to the consequences of Covid-19, partly because of anticipated complications in data gathering, it was decided that the initial participant of the questionnaire was encouraged to refer to a person of choice – someone they frequently watch television with – while filling out the survey individually. This would also ease the limitation brought forward by dyadic research, which describes that it is hard to control for dominating answers by one person in the dyad (Jameson,

2004). Jameson (2004) used the presumption that the responsibility for making a final decision lies with one member in a group, as the groundwork for his studies on group recommender systems. By letting only one person in the dyad fill out the questionnaire, we partly built on this argument, assuming that this person is in this case acting as the final decision maker of the dyad. Physically contacting someone was preferable, but as imagined, as a result of Covid-19 measures, this could bring along some complications in finding the right target audience. Hence, it was also encouraged to reach out digitally. In case of non-response, the participant was asked to proceed with the survey using input based on what they *think* the other person would like to watch. Since people were free in choosing which television partner to refer to, it was assumed a discussion about viewing interests between the two partners had likely emerged before. Moreover, co-viewing is a common activity, and is often enjoyed with someone that is relatively close to you (Chorianopoulos, 2007), thus gathering some standard demographical information about the other person (age, gender, and highest educational level) did not seem to burden the progress of the questionnaire.

The online survey tool Qualtrics was used to create the survey, since this software allowed us to pre-enter the various flows that will emerge from the selection of genre preferences. With regards to the distribution of the survey, this raises some complications. For instance, at the moment of receiving the invitation to participate, the people reached may not be with or talking to a suitable partner to enter the questionnaire with. What's more, both individuals need to be willing to take part in the survey and have the time to do so. Two rounds of demographic questions have been added to the survey to increase the likelihood that the survey will be filled out by two different individuals, or at least by taking another individual in mind.

The phases for a group recommendation process, as outlined by Jameson (2004), have been used as a reference to represent the continuation through the questionnaire as closely tied to reality as possible. Jameson (2004) has identified the following steps on how most group recommender systems go about reaching their conclusive suggestions. Firstly, the members of the group (or in our case, the dyad) specify their viewing preferences. Then, based on this information the system generates recommendations, which will be presented to the members (see the screenshots of our interfaces in appendix A). Lastly, the members engage in a decisive process about which suggestion to accept (if any). In our research, the watching intention of the three presented options will be used as a way to assess the usefulness of the system, complimented with questions about intentions for further usage and their perception on novelty, diversity, and accuracy of the recommendations.

About NPO (Start)

This research is written with the help of the Dutch public broadcaster NPO, which had great

benefits for the reliability of this study, since the ability to build on an existing platform and their content painted a more realistic picture to the participants of our study. NPO Start is the online streaming platform of the Dutch public broadcaster (NPO). The mission of the NPO is to be a public broadcaster that "connects and enriches the Dutch public with programmes that inform, inspire and entertain", showcasing content that evokes curiosity, mutual understanding and open-mindedness, while being loose from political or commercial ties (NPO, n.d.). The NPO is an administrative body functioning as an umbrella organization that provides room for content of several independent licensed broadcasting associations.

According to the NPO representative that is guiding this research, NPO's current (individual) personal recommender system consists of collaborative based filtering, for which user activity is used as input. In specific, they look at the percentage of streams that have been watched in full as an indication of content that matches the interests of a user. As the NPO explained, the first experiments using metadata (textual cues attached to a programme) have been developed, but not yet put to use for the public. Moreover, the NPO tries to stimulate people to watch a wider variety of content, such as television shows that carry a high public value. These shows have been rated as such by a panel on various items, including diversity. As argued before, television viewing is a highly sociable activity mainly enjoyed by families. Since the NPO takes responsibility in serving the Dutch population as a whole, they are ought to provide content for all niches and settings that people may watch television in together. Hence, it may be highly beneficial to them to employ a group recommender system to enhance the acceleration of their mission.

For better and easier understanding of the questions by the participants, therewith aiming to improve the completion ratio, the survey has been created in the Dutch language. This ties in well with the target audience of the NPO, as the NPO offers a vast amount of their television programmes in Dutch. For each of the genre combinations, several existing shows aired by the NPO were chosen. Tintarev (2017) pointed out that item placement and grouping of certain items together can have a major influence on the perception of diversity of the given selection, based on recall effects (for instance, the first and last item are remembered the easiest). A total of three recommendations per screen were provided to avoid placement complications, but still give the user the idea of choice (Tintarev, 2017). Based on the suggestions of the NPO Start application on the 22nd of April 2020, a selection was made from the shows under the headings on the homepage, being 'Populair' [popular], 'Nieuws & Actualiteiten' [news & current events], and 'Uitgelicht' [Highlighted], as well as the top results under the tab 'Programma's [TV shows]. A confidential document from the NPO containing so-called ccc-codes for identifying genres, was used to specify which programme belonged to what genre and, in case of doubt, if they would be suitable to be placed among the recommendations that combine two genres. For instance, if two people picked

the genres drama series and entertainment, it would make sense to put a series written and performed by two comedians in this combined genre category – which was named humorous drama. It is important to note that all television shows were chosen using the application without being logged in to a (paid) NPO Plus account, which means that no personal recommender algorithm was active that could have potentially influenced the programme recommendations based on the watching history of the researcher. According to the NPO representative, automated personal suggestions are only provided to those users who have an active NPO Plus account, as this provides the NPO with the ability to track their user activity.

Before the actual data collection of our quantitative survey research, a quick test among the same sample (N=6) was conducted to see whether the proposed genre-TV show combinations were supported, and not solely a reflection of personal interest. As the order of TV shows in the recommendations is not essential to the purpose of this research, a small sample was found to be sufficient. All 18 chosen TV shows were listed, as well as the six genre categories or combinations of those categories. The participants were then asked to place exactly three TV shows in each of the genre categories, creating a perfectly even distribution among the categories. One was encouraged to move programmes around as much as they would like, as well as to refer back to the descriptions of the TV shows written by the NPO, which were also provided.

Few changes were made accordingly and the TV shows with the strongest connection (that is, the least disagreement) to a particular genre were placed first in the list of recommendations. TV shows with a weaker connection were placed either second or third, or moved around to form a slightly stronger connection with another genre. For instance, 'De slimste mens' is a knowledge quiz in which certain phenomena or answers are supported by explanations by a historian. The show was initially placed under the genre topical interest, however, as was indicated by the majority, later moved to entertainment. Most programme-genre combinations proved successful and were left as such.

Structure of the questionnaire

Informed consent was asked at the beginning of the online questionnaire by outlining several statements on the intentions of this research while making sure that the requirements for conducting proper ethical research were met. These include, among others, a note on the fact that participation is voluntary, and the right to quit at any point, or to skip questions. Furthermore, the lack of risks associated with participation, confidential treatment of personal data, anonymity of data collection, and sole academic purpose of data gathering were highlighted. There was no need for a cover story and/or mock questions to mask the purpose of the study, so the goal of this study was touched upon briefly. Completing the survey generally took about five to ten minutes. After

reading the statements outlined, the participant gave consent by clicking the option that these terms were fully read and understood, and he/she agreed to participate in the research. The full questionnaire can be found in Appendix B.

Afterwards, some demographical questions were asked, as well as the participant's familiarity with online streaming platforms. Here, the participants were requested to check those boxes of the platforms they used before (a list of the most common ones in the Netherlands was provided). Proceeding with social relationship quality, the respondent was asked to specify the relationship he/she and the person of choice have towards one another (e.g. family, roommates, romantic relationship, friends). Besides that, he/she was asked to answer several statements on how close they would rate their relationship, similarity in their ways of thinking, and liking of the other.

Following the question on what the respondent would like to watch (genre preference, pick one out of three), the same question was asked for the person of choice. A visual image was accompanying this question, in order to give respondents a better understanding of what the NPO Start interface could look like, if this mock-up GRS tool were turned into an actual feature of NPO Start (for interface design, see Appendix A). For this, the current interface was 'updated' with a feature for "Samen TV kijken" [shared television watching], to foster people's ability to imagine a setting where one was about to decide on a television programme to watch together. Here the participant was again encouraged to reach out to someone if they had not done so already. Also, this was the point in the survey at which the demographical questions about the other person were asked, so all questions involving someone else were presented at once. This way, the participant would not have to bother someone else for the entire duration of the questionnaire, but just a small part, hopefully contributing positively to the completion rate of the survey.

Furthermore, the Qualtrics flows have been designed in a way that each combination of genre preferences (AA, AB, AC, BB, BC, and CC) leads to the desired screen showing three recommendations that fall within that genre (combination). Respondents were encouraged to inspect the screen with recommendations before continuing to the next questions and were in all cases able to click the back button, so they could be more certain of their answers if needed. The last questions evolved around the perceived usefulness of the GRS' selection of recommended videos (in terms of accuracy, novelty, and diversity), their intentions for further usage of the system, and separate watching intentions for either three of the recommendations. Lastly, a manipulation check was presented, as well as a box to leave any additional comments.

3.3 Data collection

This research contributes to a bigger study for the NPO to assess whether the use of GRS is

advantageous towards the (co-)viewing experience of their users. Hence, the sample for this research will be similar to that of the NPO Start audience, as the survey will be held among actual or potential users of the application. The target audience will consist of Dutch-speaking individuals who often engage in co-viewing behaviour either online or offline, and who are, for ethical reasoning, over the age of 18. Affinity with the Dutch language is a requirement resulting from the fact that the application of NPO Start – as well as most of its content – is available in Dutch only. Respondents will need to have access to an internet connection in order to fill out the online questionnaire, hence access to online streaming platforms could be logically derived. Familiarity with those services will be inquired by means of the questionnaire.

Ideally, the questionnaire would have been spread with the help of NPO through their community platforms, social media and the like. This would have made the sample align perfectly with the target audience of the study and would have lowered the methodological complications other sampling methods can carry. However, since the NPO was not responsive to this request, alternative sampling techniques were employed in order to reach our sample. Due to Covid-19, some sampling methods such as selective sampling were excluded from the range of options as well.

The sampling technique used to spread the questionnaire among out target audience was a combination of snowball and purposive sampling. This method allowed for a starting point among people that would fit the target group and feathering out to people that formed qualified candidates in their personal networks (Matthews & Ross, 2010). Probability sampling was not feasible in this case, since our target audience was too broad to be able to identify each individual member. Snowball sampling by the first 'generation' of participants as well as social media were used to gain access to a bigger community and variety of backgrounds of people located all over the country. Sue and Ritter (2007) argue that snowball sampling is a convenient method to make the sample more representative by relying on the network of the in-groups. Higher variation of the sample will be reached as both direct peers and outsiders will be able to share their perspectives, which are likely to differ in demographics such as (financial) background, alpha/beta studies, and geographical location (Sue & Ritter, 2007). Here, the social media networks from first-approached participants proved great tools in order to reach those audiences. Baltar and Brunet (2012) identified several advantages of what they named virtual snowball sampling. Few of these advantages include increased attractiveness to participate and thus higher response rate, time efficient data collection, and higher quality of data collected. Especially Facebook proved useful here, as its algorithm allows posts to spread relatively easily to audiences otherwise not reached, based on user interaction with the post (such as commenting and sharing; Baltar & Brunet, 2012). As a result, the reach of the survey link was largely out of the researcher's hands, making it almost

impossible to send out reminder messages as well as to indicate how many people the survey had reached.

3.4 Sample

For quantitative survey research, the methodological guidelines by Janssen and Verboord (2017) indicates that aiming between 150 and 250 responses would provide a solid basis to perform proper statistical analysis. By means of the sampling technique a total of 217 people were reached, who had all started their participation in the questionnaire. After cleaning the data, it appeared that 182 out of the 217 participants had completed the survey in full. This results in a completion rate of 83.9%. Our data was collected between April 23rd and May 16th, 2020.

Since the study has some resemblance with an experimental design in terms of the formation of different groups and its subsequent flows, it was decided to include a manipulation check. For this manipulation check, participants were asked to check the box of the third recommendation they had just seen/for which they had indicated their watching intention. Six options were provided, equivalent to the third recommendations as they were shown across the six different genre combinations. This check was placed to assess the attentiveness of participants while filling out the questionnaire, and filtering out those who did not answer correctly, will likely improve the care with which the answers were given. Besides that, a manipulation check is also beneficial for easier identification of what group each participant belonged to – if they answered correctly – in the later phase of data analysis.

Of the 182 participants that had reached and answered the final question – the manipulation check – 160 participants in turn had passed this check and 22 participants wrongly indicated the programme they had been shown before. All respondents were at least 18 years old; the survey flow directed those who were not to the end of the survey immediately. We will refer to those that finished the survey and passed the manipulation check as the valid sample for the remainder of this study. Hence, N = 160 responses were included in further analysis.

For the total valid sample, the observed age range was between 19 and 71 years old (*M* = 26.46, *SD* = 10.23). In total, 36.3% of the respondents identified as male, and 63.7% as female. There were no participants indicating ‘other’ or ‘prefer not to say’, leaving us with a binary variable which can be included in the regression analysis later on. Generally, the sample was quite well-educated, with ‘WO Bachelor’ and ‘WO Master’ being the most prominent (35.0% and 33.8% respectively). Other frequently named levels of highest education included ‘HBO’ (University of applied sciences; 18.8%), followed by ‘secondary education’ (7.5%), and ‘MBO’ (vocational education; 3.8%). The remaining 1.2% indicated their educational background either as ‘LBO’ (lower vocational education) or ‘PHD, MBA, or similar’.

We have also asked some demographical information about the person of choice, with whom the respondent would frequently engage in co-viewing behaviour. It was noticeable that viewing-partners were generally slightly older than the respondents in our survey; the observed age range was between 18 and 71 years old (*M* = 31.08, *SD* = 14.47). Of this group, the gender distribution was slightly more equal: 54.4% identified as male, and 45.0% as female. One person chose 'prefer not to say'. Similarly, the viewing-partners were generally quite well-educated too, with 'WO Master and 'WO Bachelor' again forming the biggest groups (32.5% and 26.3% respectively). Other frequently named levels of highest education included 'HBO' (University of applied sciences; 15.0%), followed by 'MBO' (vocational education; 11.9%) and 'secondary education' (9.4%). The remaining 5.0% indicated their educational background either as 'LBO' (lower vocational education) or 'PHD, MBA, or similar'.

3.5 Measurements

Five main variables were used throughout this study to test our hypotheses and calculate a potential mediating effect. These variables will be listed and discussed below, and if relevant the reliability of the scales will be presented. Lastly, a justification for one of the control variables is given.

Match in viewing interests of dyads

The first variable in this study is the match in viewing interests of dyads. This variable is constructed from the combined answer of two people's genre preferences. By means of the questionnaire the participant was asked to indicate his/her own genre preference, as well as to refer to a person of choice to either let them indicate what genre they would want to watch, or the participant could make an educated guess for a person he/she frequently watches television with. Respondents were asked to pick one option out of three. Each genre has been given a letter; A equals entertainment, B equals drama series, and C equals news & current events. Based on the answers to both questions, combinations of those letters could be made, as shown in Table 3.1. Although the question was asked in Dutch, the English genre names as well as their corresponding group and number of participants are provided.

There are six conditions linked to the combination of genre preferences (AA, AB, AC, BB, BC, and CC). The dyads with different preferences (AB, AC, and BC) have been combined into one umbrella group and recoded into 0. The same was done for the dyads with similar preferences (AA, BB, and CC), which were recoded into 1. We argue that the dyads with similar preferences will differ in their watching intention of the recommended videos, insofar they essentially would not benefit as much from the recommender system than would the dyads with combined preferences. Since

their preferences are more alike, we assume the shown recommendations will form a better match with the actual preferences of the individuals in the dyad, which will be analysed by means of H1. In total, the group with similar preferences consisted of 100 dyads, whereas the group with different preferences included 60 dyads.

Table 3.1. *Match in viewing interests of dyads explained*

Combination of genre preferences	*English genre names*	*Number of participants per group*	*Recoded into binary numbers; 0: N= 60, 1: N= 100*
AA	Entertainment	37	1
AB	Humorous drama (series)	30	0
AC	Satire	14	0
BB	Drama series	49	1
BC	Topical interest	16	0
CC	News & current events	14	1

Social relationship quality

Our second variable, social relationship quality, is based on a pre-existing scale from Hennig-Thurau, Marchand, and Marx (2012) which has been copied almost in its entirety, only inserting those items that addressed the relationship between the self and the movie-partner. Social relationship quality was measured with six items, three items on liking of the co-viewer (being "I like my movie-partner very much as a person.", "I think my movie-partner is a good friend.", and "I get along well with my movie-partner."), and three items on the similarity of the self and the co-viewer (being "My movie-partner and I are similar in terms of our outlook, perspective, and values.", "My movie-partner and I see things in much the same way.", and "My movie-partner and I are alike in a number of areas."). Respondents were asked to indicate whether the statement applied to them on a Likert scale from 1, Strongly disagree to 5, Strongly agree.

The items were entered into factor analysis using Principal Components extraction with Varimax rotation based on Eigenvalues (> 1.00), *KMO* = .73, $\chi 2$ (*N* = 160, 21) = 272.42, $p < .001$. The resultant model explained 59.2% of the variance in social relationship quality. Factor loadings of individual items onto the two factors found are presented in Table 3.2. The factors found were in accordance with liking and similarity, as would be expected from the original scale. Initially, a visual depiction of closeness to the other by means of two overlapping circles, as they are also frequently implemented in other studies, were included in the questionnaire as well (see: Aron, Aron, &

Smollan, 1992). However, as appeared from reliability analysis, leaving out the latter contributed to a higher Cronbach's alpha score on one of the factors, hence these circles were not taken into further consideration. In order to appeal better to our target audience, all scales were translated into Dutch, which may have affected the reliability of the scale slightly. The first factor's Cronbach's alpha was .69 after deleting the circles, and .69 for the second factor, meaning the measurements are slightly below the reliable threshold. However, according to Shelby (2011), researchers sometimes consider a lenient criterium between .65 and .70 to form an 'adequate' scale. Shelby (2011) on the other hand, critiqued the use of Cronbach's alpha of being misleading and often being interpreted too simplistically, especially among scales with few questions. Instead, he suggested to consider a corrected item-total correlation above .40 as a criterium for scale reliability, in addition to the Cronbach's alpha score. Although Hennig-Thurau, Marchand, and Marx (2012) obtained a higher Cronbach's alpha score for each factor than was the case in the present study, we still looked at our corrected item-total correlations. For the first factor, corrected item-total correlations of .55, .43, and .51 were found, hence indicating an acceptable scale. For the second factor, the scores read .45, .56, and .53, also providing an acceptable scale. The factor scores were calculated by averaging respondents' score on the three items each. For the first factor the scores ranged from 2.33 to 5.00 (*M* = 4.50, *SD* = 0.54). The scores for the second factor ranged from 2.00 to 5.00, (*M* = 3.68, *SD* = 0.63). An overview of the means and standard deviations of the continuous variables can be found in Table 3.3.

Table 3.2. *Factor and reliability analyses for scales for social relationship quality (N = 160)*

Item	*Liking*	*Similarity*
I like my movie-partner very much as a person.	.79	-
I think my movie-partner is a good friend.	.68	-
I get along well with my movie-partner.	.77	-
My movie-partner and I are similar in terms of our outlook, perspective, and values.	-	.80
My movie-partner and I see things in much the same way.	-	.79
My movie-partner and I are alike in a number of areas.	-	.72
M	4.50	3.68
SD	0.54	0.63
Cronbach's α	.69	.69

Perceived relevance of GRS

Next, perceived relevance of the Group Recommender System (GRS) was based on a pre-existing scale as well, which was copied from Pu, Chen, and Hu (2011). Pu, Chen, and Hu (2011) referred to the scale items as the construct for use intentions. The relevance of a recommender system can be made operationalizable by looking at people's intentions to make use of the system. In our case, if people have a high tendency to use the GRS, it indicates that the system carries a high relevance – otherwise they would most certainly opt for an alternative manner of choice facilitating. This idea was supported by Bennett and Lanning (2007), and Gomez-Uribe and Hunt (2015), who stressed the importance of being both relevant, transparent, and easy to navigate in order to minimize chances of abandoning the platform. Our questionnaire was constructed in such a way that participation was automatically connected to making use of the GRS mock-up, thus relevance was measured by means of respondents' further use intentions of the system. This also gave participants the chance to familiarize themselves with the mock-up before they were asked to assess its relevance.

Perceived relevance of the GRS, or further use intentions, was measured using three items, being "I will use this recommender again.", "I will use this recommender frequently.", and "I will tell my friends about this recommender.", which were also translated into Dutch. Respondents were asked to indicate whether the statement applied to them on a Likert scale from 1, Strongly disagree to 5, Strongly agree. The items were entered into factor analysis using Principal Components extraction with Varimax rotation based on Eigenvalues (> 1.00), *KMO* = .60, $\chi 2$ (*N* = 160, 3) = 117.47, $p < .001$. The resultant model explained 64.2% of the variance in perceived relevance of the GRS. As expected, only one factor was found. Reliability analysis showed a Cronbach's alpha score of .71, meaning it was already a reliable measurement. A major increase in the Cronbach's alpha was found after deleting the last item ($\alpha = .80$), hence a new variable was created using the average scores on the remaining items, which were used for further analysis. These scores ranged from 1.00 to 5.00 (*M* = 3.19, *SD* = 0.82).

Perceived usefulness of GRS' selection of recommended videos

The mediator variable in this study is the perceived usefulness of the Group Recommender System's selection of recommended videos, which was based on a pre-existing scale too. The measurement was adapted from Pu, Chen, and Hu (2011) by combining two separate questions, and one constructs consisting of two questions into one measurement:

1. Recommendation Accuracy

   "The items recommended to me matched my interests."
2. Recommendation Novelty

"The items recommended to me are novel."

"The recommender system helped me discover new products."

3. Recommendation Diversity

   "The items recommended to me are diverse."

This was in accordance with the argument of Tintarev (2017), who stated that usefulness is an interplay between user interest, novelty, and diversity. Again, respondents were asked to indicate whether the statement applied to them on a Likert scale from 1, Strongly disagree to 5, Strongly agree.

Unfortunately, factor analysis showed that these four items did not make a reliable measurement if they were combined: Using Principal Components extraction with Varimax rotation based on Eigenvalues (> 1.00), resulted in *KMO* = .43, $\chi 2$ (*N* = 160, 6) = 55.97, *p* < .001. Based on the construct from its original source, it was decided to perform reliability analysis on the two items constructing novelty, resulting in a Cronbach's alpha score of .62. We are aware that this Cronbach's alpha score is still not ideal, yet the corrected item-total correlations for novelty are both .45. According to Shelby (2011) this would indicate an acceptable scale. Following the rationale by Shelby (2011) leads us to believe it would cause no further harm to the reliability of this study to continue with this combined variable. Furthermore, the previous use by Pu, Chen, and Hu (2011), and given the academic relevance of investigating the interplay between accuracy, novelty and diversity, we decided to continue our analysis using these questions as separate items constructing our measurement for usefulness. The new variable based on the average scores for novelty ranged from 1.00 to 5.00 (*M* = 2.73, *SD* = 1.00). The scores for the other two variables ranged from 1.00 to 5.00 as well. Both the accuracy and the diversity item had a mean score of 3.31 and a standard deviation of 0.88.

Watching intention (of the GRS' selection of recommended videos)

The last main variable for analysis is the watching intention of the GRS' selection of recommended videos. As Pu, Chen, and Hu (2011) pointed out, measuring intention can be used to measure behaviour as the former is said to be a strong predictor of the latter. The measurement was adapted from a single-question construct by Pu, Chen, and Hu (2011), replacing the word 'buy' with 'watch', resulting in the item "I would watch the [first, second, third] item recommended, given the opportunity." The original question was used in relation to a mock-up recommender system, which perfectly applied to our study as well. As our interface showed a total of three recommendations per genre-combination, the question was adapting to ask for watching intention for each of the recommendations specially, since the influence of programme preference, mood, etcetera may still

be present within genres.

Furthermore, respondents were asked to indicate whether the statement applied to them on a Likert scale from 1, Strongly disagree to 5, Strongly agree, also including option 6, I have seen this programme before. Familiarity with the recommendations shown, especially those that were consumed previously, could be of major influence on our results as it can be a source for polarizing effects. For instance, people could argue not wanting to watch a show again because they have already seen it, or because they hated it. On the other hand, some people may want to watch again because they loved it. The option for respondents to indicate previous engagement with the programmes recommended – and then deleting those answers from the analysis - takes away potential polarizing effects that may transform the underlying relationship.

For the sake of better and easier interpretation of our results, the separate scores on watching intention (after deleting 6) were combined into one mean score on watching intention of the set of recommendations. So far, no theoretical argument could be made expecting significantly different outcomes per watching intention for each of the three recommendations. A combined watching intention also allows for filtering out extreme (dis)satisfaction rates of the recommendations in comparison to the rest, as the scores are now averaged with the scores for watching intention of two other viewing suggestions. This could help to make sense of the overall selection of recommendations, and thus the effectiveness of the GRS as a whole, more easily. The questions do not belong to one scale, and were not intended to measure the same concept. In fact, the three questions were answered based on three different inputs (watching suggestions), hence no factor- and/or reliability analysis was performed for this measurement. The scores for the averaged watching intention (without 6) ranged from 1.00 to 5.00 (*M* = 2.98, *SD* = 0.90).

Table 3.3. *Means and standard deviations of the continuous variables*

	M	*SD*
Social Relationship quality		
Liking	4.50	0.54
Similarity	3.68	0.63
Perceived relevance of GRS	3.19	0.82
Perceived usefulness		
Accuracy	3.31	0.88
Novelty	2.73	1.00
Diversity	3.31	0.88
Watching intention	2.98	0.90

Familiarity with online streaming platforms

One of the control variables used in mediation analysis, aside from demographic variables, is familiarity with online streaming platforms. This question asked which online streaming platforms the respondents had used before, of several popular online streaming platforms in the Netherlands that were listed (e.g. Netflix, HBO, Videoland). An option to indicate 'Other' was also provided, with a textbox to elaborate on this answer. It was encouraged to check multiple boxes. Therefore, the count of checked boxes could be summed up into a total count for familiarity with online streaming platforms, which made it a continuous variable with scores ranging from 1.00 to 10.00 (*M* = 3.45 and *SD* = 1.53). Here the assumption is made that more checked boxes means a higher familiarity with online streaming platforms, compared to those who checked little to no boxes, or entered "none" in the Other category. It could be argued that familiarity with online streaming platforms indicates a higher usage of similar platforms, and thus potentially a higher benefit of a group recommender system. On the other hand, higher familiarity may also indicate a more critical approach to our mock-up GRS. Since a theoretical basis for this thought is not yet established, this measurement was added as a control variable to see whether it would yield interesting results that could be taking into consideration for future research.

4. Results

In the following chapter the collected data from our questionnaire was analysed and reported accordingly. We have identified whether the expected relationships between the variables of interest were present, and thus in accordance with our hypotheses. Based on pre-existing literature, usefulness of the recommender system will be assessed for its mediating power. Hayes parallel multiple mediation analysis was conducted to assess the associations between the variables constituting our last three hypotheses – perceived relevance, perceived usefulness and watching intention.

4.1 Hypothesis testing

After establishing the reliability of the variables, we can proceed with testing our hypotheses. The data analysis was conducted using the IBM SPSS Statistics software, version 25. The use of continuous and/or categorical variables determined which analytical tests were performed. See Figure 2 for an overview of the standardized regression coefficients and significance for all hypotheses.

Match in viewing interests

Our first hypothesis read as follows: Compared to those dyads who entered different viewing interests, dyads who entered similar viewing interests in the GRS will indicate a higher watching intention of the selection of recommended videos. The categorical independent variable match of viewing interests consists of two values indicating either a combination of different viewing interests (e.g. AB, recoded into 0) or a combination of similar viewing interests (e.g. AA, recoded into 1). The dependent variable, combined watching intention of the GRS' recommended videos, is a continuous variable, hence an independent-samples T-test has been conducted. This analysis compares the two groups on their mean scores for watching intention of the recommendations. It was not possible to add control variables to this analysis. There was no significant difference in scores for dyads with similar viewing interests (M = 2.96, SD = 0.97) and dyads with different viewing interests (M = 2.99, SD = 0.79); t (143) = 0.19, p = .848. In contrast to our expectations, the mean scores for watching intention were almost alike for both groups, thus rejecting H1.

Social relationship quality and watching intention

Our second hypothesis predicted that a higher perceived social relationship quality would be of positive influence on the watching intention of the selection of recommended videos as outlined by the GRS. Both factors that were found to construct the scale for social relationship quality (SRQ) –

liking and similarity – are continuous variables, as well as the dependent variable combined watching intention. A multiple regression analysis has been conducted, to determine if the variables are positively related. The items for age, gender, and familiarity with online streaming platforms were added as control variables. The variable age was already continuous, but in order to be able to use the variable for gender, the two present values were recoded into 0 (male) and 1 (female). Familiarity with online streaming platforms was recoded into a count of the number of platforms checked, as explained in the methodology, hence forming a continuous variable as well. For the remainder of this chapter we will refer to this variable in our tables using the shortened version 'Online streaming'. The variable on combined watching intention of the recommendations was entered as the criterium. Both factors for SRQ, and the three control variables predicted 4.2% of the variance on watching intention (Predicted R^2). Since the model proved non-significant ($F(5, 139) = 1.22$, $p = .304$), we can conclude that H2 has to be rejected. Based on the standardized regression coefficients (β -values), we can say that the pattern in the data was in line with our hypothesis; respondents who indicated a higher social relationship quality with their person of choice also showed a higher watching intention (see Table 4.1). Yet, even with the inclusion of control variables, the predicted variation remains very low.

Table 4.1. *Model summary of multiple regression analysis for Social Relationship Quality (SRQ), N = 145*

Predictor variables	β	*p*	*F*	*df*	*p*	R^2
Overall model			1.22	5, 139	.304	.04
SRQ liking	.02	.833				
SRQ similarity	.05	.565				
Age	-.12	.163				
Gender	-.05	.589				
Online streaming	.11	.204				

Note: R^2 represents the Predicted R-squared score. The dependent variable for this regression was watching intention.

Perceived usefulness and watching intention

Prior to our mediation analysis, hypotheses three until five will be tested for their potential relationship, and if they occur in the anticipated direction. All the variables in these analyses are continuous, hence we have conducted several multiple regression analyses. The third hypothesis constituted of the following: Higher perceived usefulness of the GRS' selection of recommended videos will increase user's watching intention. Perceived usefulness is measured using three

separate variables: accuracy, novelty, and diversity. These three items were added as the predictors, alongside the three control variables age, gender, and familiarity with online streaming platforms. Again, the variable on combined watching intention was entered as the criterium. Accuracy, novelty, diversity, and the three control variables predicted 35.2% of the variance on watching intention (R^2). This result was significant; $F(6, 138) = 12.48$, $p < .001$. Both novelty and accuracy were found to be significant predictors for watching intention, while diversity nor the three control variables age, gender, and familiarity with online streaming platforms were significant (see Table 4.2). In conclusion, two out of the three items constructing perceived usefulness proved significant, therefore we can partially accept H3. Despite our expectations, diversity of the recommendations was no significant predictor for the variance on watching intention.

Table 4.2. *Model summary of multiple regression analysis for perceived usefulness, N = 145*

Predictor variables	β	*p*	*F*	*df*	*p*	R^2
Overall model			12.48	6, 138	<.001	.35
Accuracy	.53	<.001				
Novelty	.30	<.001				
Diversity	-.09	.200				
Age	-.00	.988				
Gender	-.03	.635				
Online streaming	.08	.262				

Note: R^2 represents the Predicted R-squared score. The dependent variable for this regression was watching intention.

Perceived relevance and perceived usefulness

Our fourth hypothesis predicts the positive relation between perceived relevance of a GRS (by means of its further use intentions) and the perception of usefulness of the GRS' selection of recommended videos. The scale for further use intentions consists of one variable, however, as perceived usefulness is consistent of three variables, we will have to conduct a multiple regression analysis thrice. For this reason, a Bonferroni correction is necessary, as our model is based on three dependent variables. A Bonferroni correction adjusts the probability (*p*) value by dividing the critical value by number of tests performed (Armstrong, 2014). In our case instead of the usual .050 value, a value of .017 will be adopted for testing this hypothesis. Age, gender, and familiarity with online streaming platforms were added as control variables again.

First, accuracy was entered as a criterium, with further use intentions, and the three control variables predicting 15.1% of the variance on accuracy (R^2). This result was significant; $F(4, 155) = 6.88$, $p < .001$. Only further use intention was found to be a significant predictor for accuracy (see Table 4.3). The results for the three control variables age, gender, and familiarity with online streaming platforms age were not significant.

Second, novelty was entered as a criterium, with further use intentions, and the three control variables predicting 5.4% of the variance on novelty (R^2). This result was not significant; $F(4, 155) = 2.19$, $p = .072$. As a result of the Bonferroni correction, further use intention was now regarded non-significant in predicting novelty (see Table 4.3). The results for the three control variables age, gender, and familiarity with online streaming platforms age were not significant too.

Third, diversity was entered as a criterium, with further use intentions, and the two control variables predicting 7.3% of the variance on diversity (R^2). This result was not significant; $F(4, 155) = 3.04$, $p = .019$. However, looking at the coefficients, the variable on further use intentions does prove to be a significant predictor for diversity (see Table 4.3). The results for the three control variables age, gender, and familiarity with online streaming platforms age were not significant.

To sum up, since two out of three analysis proved significant results for the association of further use intentions with the three items constructing the perceived usefulness measurement, we can partially accept H4. Contrary to our expectations and our findings before, the association of further use intentions with novelty did not yield any significant results here.

Table 4.3. *Model summary of multiple regression analysis for perceived relevance (further use intentions), N = 145*

Dependent variables	*Predictor variables*	β	*p*	*F*	*df*	*p*	R^2
Accuracy	Overall model			6.88	4, 155	<.001	.15
	Further use intentions	.35	<.001				
	Age	-.04	.630				
	Gender	-0.7	342				
	Online streaming	.10	.173				
Novelty	Overall model			2.19	4, 155	.072	.05
	Further use intentions	.18	.031				
	Age	-.06	.502				
	Gender	.09	.240				
	Online streaming	-.08	.301				

Diversity	Overall model			3.04	4, 155	.019	.07
	Further use intentions	.27	.001				
	Age	.11	.175				
	Gender	-.07	.343				
	Online streaming	.05	.570				

Note: R^2 represents the Predicted R-squared score. The dependent variables for the regressions were the three items on perceived usefulness. A Bonferroni correction with a probability level of .017 was applied.

Perceived relevance and watching intention

Our fifth hypothesis was partially created by the previous two hypotheses; higher perceived relevance of a GRS (further use intentions) of a GRS will increase the watching intention of the selection of recommended videos. The scale for further use intentions consisted of one variable, which was added as the predictor, alongside the three control variables age, gender and familiarity with online streaming platforms. The variable on combined watching intention was entered as the criterium. Further use intentions and the three control variables predicted 13.0% of the variance on watching intention (R^2). This result was significant; $F(4, 140) = 5.23$, $p = .001$. Only further use intention was found to be significant predictor for watching intention (see Table 4.4). The three control variables age, gender, and familiarity with online streaming platforms were not. Thus, we can conclude that perceived relevance of a GRS (further use intentions) positively contributes to the watching intention of the recommended videos in our GRS' selection. As this association was significant, we can accept H5.

Table 4.4. *Model summary of multiple regression analysis for perceived relevance (further use intentions), N = 145*

Predictor variables	β	*p*	*F*	*df*	*p*	R^2
Overall model			5.23	4, 140	.001	.13
Further use intentions	.32	<.001				
Age	-.03	.710				
Gender	-.05	.558				
Online streaming	.08	.309				

Note: R^2 represents the Predicted R-squared score. The dependent variable for this regression was watching intention.

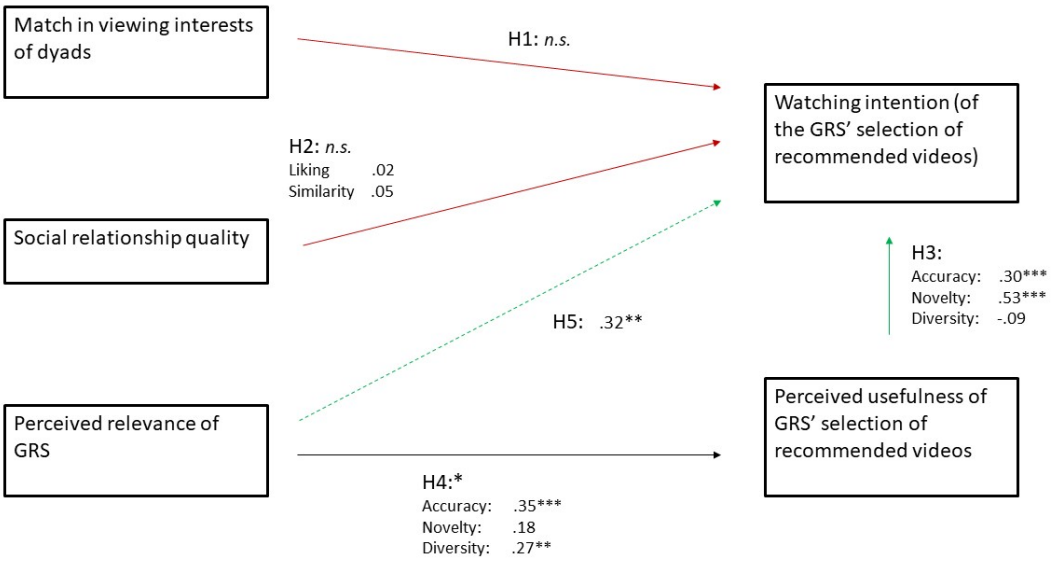


Figure 2. *Conceptual framework with standardized regression coefficients.*
*Note: * A Bonferroni correction with a probability level of .017 was applied. Significance: ** p = .001, *** p < .001.*

4.2 Mediation analysis

In order to find out whether the association between perceived relevance of a GRS (further use intentions) and watching intention of the GRS' selection of recommended videos is mediated by the perceived usefulness of this selection, a mediation analysis has been conducted. To do so, we have used Hayes' model on parallel multiple mediation (see: Hayes, 2018). First, the beta values of the relationships and their significance have been established, after which bootstrapping told us the significance of the (total) indirect and direct effects. This gives us insight into the mediating power of perceived usefulness on the association between X and Y. It was decided to use Hayes' PROCESS macro v3.5 for SPSS as the most recent software to calculate mediation among variables, using bootstrapping to test the statistical significance of the indirect effects. The three constructs for perceived usefulness – accuracy, novelty, and diversity were functioning as mediator variables. Age, gender and familiarity with online streaming platforms were added as control variables (covariates) throughout the procedure. For a conceptual diagram of the pathways see Figure 3, for a conceptual diagram including the significance of the relationships and beta weights, see Figure 4.

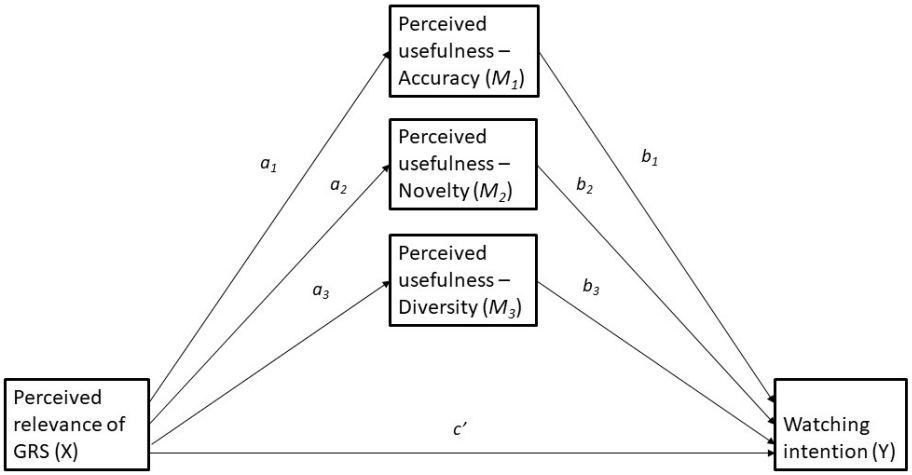


Figure 3. *Conceptual diagram of the parallel multiple mediator model, indicating the pathways between variables.*

First, as can be seen in Figure 4, higher further use intention (as the measurement for perceived relevance) was related to higher perceived accuracy of the GRS' recommendations (a_1 = .42, p < .001). The control variables age, gender, and familiarity with online streaming platforms had no significant influence on this relationship (see Figure 4 for the regression coefficients, standard errors and p-values). This model predicted 17.9% of the variance on accuracy (R^2). This result was significant; F(4, 140) = 7.62, p < .001. Subsequently, higher accuracy was related to a higher score for watching intention (b_1 = .48, p < .001). Again, none of the control variables age, gender, and familiarity with online streaming platforms had a significant influence on watching intention (see Table 4.5).

Second, higher further use intention was not significantly related to higher perceived novelty of the GRS' recommendations (a_2 = .15, p = .159), although there was a significant association between perceived novelty and watching intention (b_2 = .26, p < .001). The control variables age, gender, and familiarity with online streaming platforms had no significant influence on the former relationship; non-significance for the latter was already established (see Table 4.5). The first model predicted 3.0% of the variance on novelty (R^2). This result was non-significant; F(4, 140) = 1.10, p = .360.

Third, higher further use intention was significantly related to higher perceived diversity of the GRS' recommendations (a_3 = .26, p = .007), but there was no significant association between perceived diversity and watching intention (b_3 = -.12, p = .109). Among the control variables age, gender, and familiarity with online streaming platforms no significant influence on the former relationship was found either; non-significance for the latter was already established (see Table 4.5). The first model predicted 6.7% of the variance on diversity (R^2). This result was significant; F(4,

140) = 2.50, p = .045.

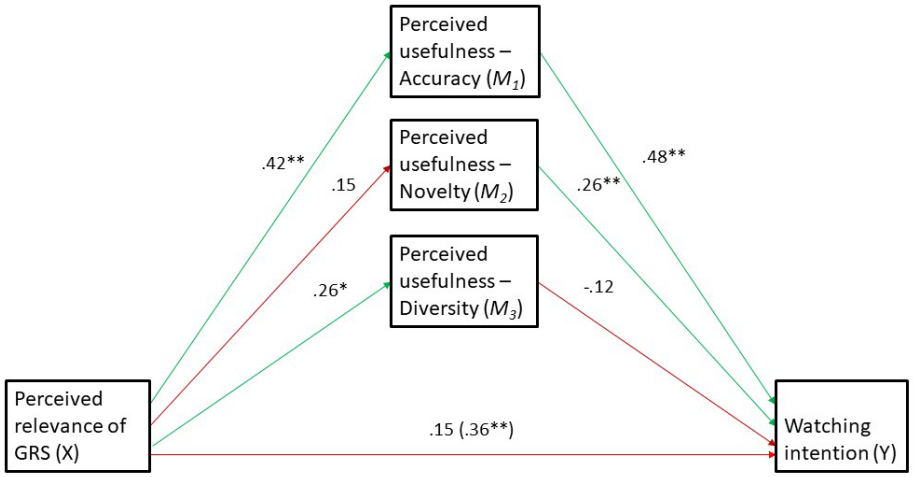


Figure 4. *Standardized regression coefficients for the relationships between perceived relevance of GRS and watching intention as mediated by the three constructs for perceived usefulness (accuracy, novelty and diversity). The standardized regression coefficient for the indirect effect between perceived relevance of GRS and watching intention is in parentheses. * p < .010, ** p < .001.*

Thus, if users indicated a higher further use intention (as the measurement for perceived relevance), this would be significantly and positively associated with accuracy and diversity of the recommendations, but only accuracy and novelty appear to have a significant positive influence on the watching intention of these recommendations. The relationship between further use intention and watching intention was mediated by accuracy of the recommendations. As Figure 4 illustrates, the standardized regression coefficient between further use intention and accuracy was statistically significant, as was the standardized regression coefficient between accuracy and watching intention. The standardized indirect effect for a_1b_1 was (.42)(.48) = .20. The significance of this indirect effect was tested using bootstrapping. A 95% bias-corrected confidence interval based on 5,000 bootstrap samples indicated that the total indirect effect was entirely above zero (0.08 to 0.37), which means that it is significant. The indirect effect through perceived accuracy, holding all other mediators constant, was entirely above zero (0.10 to 0.34), indicating the significant indirect effect for accuracy as an individual mediator. In contrast, the indirect effects through both novelty and diversity were not different than zero (−0.02 to 0.10 and −0.09 to 0.01, respectively; see Figure 4 for the beta values associated with these pathways). The coefficients, standard error estimates (S.E.) and p-values can be found in Table 4.5. All models were not subject to any predicting influence from the three control variables, which means that the results found can be solely attributed to the initial variables of interest.

Moreover, as can be seen in Figure 4, higher further use intention was found to have a significant indirect effect on watching intention ($c = .36$, $p < .001$), but no significant direct effect through all three measurements for perceived usefulness ($c' = .15$, $p = .11$). Given that the mediation fully explains the variance of Y by X, we could speak of full mediation (Baron & Kenny, 1986). However, Hayes (2018) also critiqued the use of full- and partial mediation, since there are chances that not all variables explaining the mediation are captured in this model. For instance, there could be additional explanatory elements that mediate the association between further use intention and watching intention just as much as perceived usefulness, that we are simply not measuring (Hayes, 2018). In our case, the effect of the mediation by accuracy is high enough to make the entire model significant, even though both diversity and novelty were no significant mediators on their own. Rather, for the discussion of our results, we will argue that the relationship between further use intentions (perceived relevance) on watching intention is mediated by the accuracy construct of perceived usefulness, whereas the effect of novelty and diversity is open for interpretation, as will be further elaborated on in the next chapter.

Table 4.5. *Regression coefficients, standard errors, and model summary information for the perceived usefulness parallel multiple mediator model depicted in Figure 4.*

		Consequent											
		M_1 (Accuracy)			M_2 (Novelty)			M_3 (Diversity)			Y (Watching intention)		
Antecedent		*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*
X (Relevance)	a_1	0.42	0.09	<.001	a_2 0.15	0.11	.159	a_3 0.26	0.10	.007	c' 0.15	0.09	.106
M_1 (Accuracy)		---	---	---	---	---	---	---	---	---	b_1 0.48	0.08	<.001
M_2 (Novelty)		---	---	---	---	---	---	---	---	---	b_2 0.26	0.07	<.001
M_3 (Diversity)		---	---	---	---	---	---	---	---	---	b_3 -0.12	0.07	.109
Age		-0.01	0.01	.462	-0.01	0.01	.427	0.01	0.01	.155	0.00	0.01	.668
Gender		-0.14	0.15	.354	0.13	0.17	.444	-0.16	0.15	.285	-0.08	0.13	.564
Online streaming		0.04	0.05	.457	-0.03	0.06	.654	0.05	0.05	.339	0.05	0.05	.293
Constant	i_{M1}	2.01	0.46	<.001	i_{M2} 2.56	0.52	<.001	i_{M3} 2.14	0.47	<.001	i_Y 0.40	0.50	.427
		R^2 = .18			R^2 = .03			R^2 = .07			R^2 = .36		
		$F(4, 140) = 7.62$, $p < .001$			$F(4, 140) = 1.10$, $p = .360$			$F(4, 140) = 2.50$, $p = .045$			$F(7, 137) = 11.20$, $p <.001$		

Note: Age, Gender and Online streaming refer to the control variables (covariates)

5. Discussion and conclusion

In conclusion, we were interested to what extent our mock-up version of a group recommender system can positively contribute to the co-viewing experience of its users. This association has been investigated among dyads, and by measuring the watching intention of the system's recommendations on online public television content, based on the combined viewing interest of the dyads. It was found that usefulness of the recommendations – in terms of accuracy – was mediating the positive relationship between further use intentions and the combined watching intention of all three recommendations. Since further use intentions was our construct to measure the relevance of such a group recommender system, we can argue that perceived relevance leads to increased intentions to engage with the content endorsed by the GRS, if these recommendations seem accurate in relation to (previous) user interests. Diversity and novelty of the recommendations did not have a mediating effect, although the overall model including all three constructs for perceived usefulness was significant. No evidence was found for a difference among viewers with a match or difference in genre preferences, or the social relationship quality of our dyads. To provide an answer to our research question, this study found that the output from our GRS was rated quite positively, and positive associations were found for both usefulness of the recommendations and watching intention and usefulness of the recommendations and further use intention of this GRS, which indicates the relevance of such a system. Although we could conclude that our GRS contributed to the co-viewing experience of its users and the decision-making process leading up to it, the design carries some limitations which will be addressed in the following paragraphs. Generally, (group) recommender systems are built around algorithms to evaluate previous user activity and propose items based on certain pre-entered criteria (such as focusing on novelty and/or diversity, limits on items that are too similar etcetera; Davidson et al., 2010; Sørensen & Hutchinson, 2017). Our GRS tool also illustrated the current debate on accuracy vs. diversity (and novelty) in recommender systems. Supported by our findings, we argue that all of these criteria are important in creating a well-functioning group recommender system, while we also point out that some of these items are harder to measure, and thus harder to capture in a good GRS effectiveness assessment model. Furthermore, we will into the theoretical and methodological explanations of our findings, also touching upon the non-significance of match in viewing interests, and social relationship quality. Along the way, we provide some suggestions for further investigation and discuss some of the practical implications of the outcomes of this study.

5.1 Discussion

Mediation analysis

Our mediation analysis showed full mediation insofar the indirect effect between further use intention and watching intention is significant, and the direct effect between the two is not. Yet, if looking at the effects of separate mediators, only the item on accuracy shows full mediation, whereas both novelty and diversity are not. In line with contemporary research on most commercially focused group recommender systems, it is indeed relevant to portray a high accuracy in their recommendations. This is arising from a business perspective, that is dealing with high competition in terms of attentions spans and tries to appeal to their audience by increasing the chances they find something relevant quickly (Gomez-Uribe & Hunt, 2015). Moreover, the user requests transparency of the recommendations in order to find the system reliable; he/she wants to know *why* these programmes were recommended to him/her (e.g. based on previous user activity; Sørensen & Hutchinson, 2017). In many cases, the easiest way to fulfil both criteria, is to stick to accuracy-based recommendation algorithms. As Vargas (2011) puts it, the objective of a recommender system is to "satisfy the seller's interest by satisfying user interests" (p. 8). In other words, a GRS that is able to reflect user interests well can be considered accurate in giving recommendations, which in turns helps the retention rates of the GRS. Zhang (2013) also concludes that in the current debate on recommender systems, over 90% is dedicated to accuracy metrics. As confirmed by this research, accurate recommendations have a positive relationship with the watching intention of the endorsed videos outlined by the GRS. We also established that the link between further use intentions and watching intention was mediated by accuracy, indicating that accuracy is also considered valuable on the long term. However, recommender systems that have a sole focus on accuracy can be at risk of endorsing *monothematic items* (Vargas, 2011). With this Vargas (2011) meant giving recommendations that can be perceived as all the same, instead of separate recommendations. Likely, the user will find these suggestions too obvious or too similar to one another to see the benefit of the system (Zhang, 2013). The same applies to systems that focus on popularity of items to create a sequence in the recommendations (that is, the most popularly rated item by others comes first). Even if the user had no previous engagement with the item, and it is in line with previous watching behaviour – thus: accurate – chances are high that due to the item's popularity, the user is already familiar with it, or could have easily found it on his/her own. This leaves little relevance for the recommender system, as they are employed to provide users with content that is harder to find, yet relevant, within a limited time span. That is why other criteria such as novelty and diversity have often been considered in the development of (G)RS (e.g. Vargas, 2011; Zhang, 2013).

Inserting novelty in recommender systems can be slightly trickier, as it builds on proper recommender system accuracy. Novelty concern those items that the user was not yet familiar with, but in reality a set of recommendations often contains both known and unknown items (Zhang, 2013). This is because a system can hardly measure what items a user has encountered before, only those items that were watched, liked, disliked etcetera, can be taken into consideration. In our research, we reconstructed this element by adding the option to indicate that one had already seen the programme presented to them. Furthermore, the system attempts to find those novel items that have a considerate match with the watching history of a user, which not always form an evident connection. On the other hand, a set of recommendations that are all novel have the risk of not being accurate anymore, and not satisfying user needs. Perhaps, this argument could explain the non-significant results for the association of further use intentions with novelty. In other words, providing novel items does not necessarily increase the intentions to further use the system if they do not accurately represent the need of the user. Oftentimes, in attempting to overcome this burden, it is assumed that less popular items have higher novelty for users (Zhang, 2013). Our results proved that users still value novelty of their recommendations and that this is a strong predictor of watching intention of the GRS' recommendations. It could be argued that based off generic recommendation structures embedded into the NPO Start application that foster widening one's watching portfolio, less popular recommendations were selected and inserted into our mock-up GRS. Following that rationale, it would seem reasonable that differences per genre (combination) existed. Overall, the mean score for novelty was 2.73, which was lower than those for accuracy and diversity (3.31 for both). This is an interesting number, since all lowest extremes (that is, all that indicated they had seen the programme before) were not inserted into this calculation. This once again illustrates that novelty cannot be measured looking solely at viewing history and/or engagement with TV shows. Due to external factors some programmes from the NPO may have received more popularity than others, influencing the novelty score of our watching suggestions.

Moreover, the criteria on diversity was not found to be a proper predictor of the watching intention of the selection of videos portrayed by the GRS. Whereas novelty is seen as an item that is similar to watching history topic-wise but contains new information, diversity aims to broaden one's perspectives and to burst one's filter bubble (Zhang, 2013). Filter bubbles are the reinstatement of one's own viewpoints, by the subconscious act of preferring content that is equal to your interests, thus carrying the risk of overlooking content that is important to contemporary understanding of society (Nagulendra & Vassileva, 2014). Public broadcasting media use a diversified media portfolio in an attempt to break through this bubble and educate their audiences on all sorts of topics. In this light, diversity may not always yield results in line with user's

expectations and will not satisfy their needs at all times. Yet, diversification of recommendations does not necessarily undermine a proper result, since fostering to think outside of the box will not be as easily accepted as those programmes that are perfectly in line within one's familiar watching preferences. Following this rationale, the non-significance results could be somewhat logically derived, but that does not mean diversity has no value to the non-commercial intentions of this GRS. Our results have shown that there is a significant positive association between further use intentions and diversity. Hence, we build from there in arguing that our respondents do value diversity when it comes to intentions to make further use of the GRS.

Another potential explanation for deviating findings for the diversity construct of perceived usefulness, could be assigned to the nature of the concept. Diversity as a concept entails that there is something to differentiate between; that is, diversity is best measured if it can compare one programme with the other and reach a conclusive diversity score on the *set* of recommendations (Silveira, Zhang, Lin, Liu, & Ma, 2019). The risk with indicating a diversity score for individual items, is that the first few have very little ground for comparison, only as the number of programmes to rate increases, will the thoroughness of answers given improve. To illustrate, accuracy and novelty of an item can be rated more easily after seeing just one, comparing the TV show to one's own watching history. For accuracy, one can seek for a match in usual programme preferences, and for novelty, one can simply ask whether the TV show is new to the user. Diversity, on the other hand, is a more complex measurement, since it is likely deviating from genre preferences (accuracy), but also novel. Additionally, diversity can also apply to in-between case comparisons, as bigger differences among the range of recommendations can also lead to a higher diversity score. Thus, in order to properly answer this question, it would be easier if a comparison between the other items in the set were possible and/or a combined diversity score for all recommendations could be given. It is definitely encouraged to explore this when duplicating this study.

In conclusion, only accuracy is fully mediating the association between further use intentions and watching intention of the recommendations by the GRS. In line with our argumentation above, we can say that in order for a GRS to be relevant and for users to actually follow up on the recommendations provided, these recommendations should be accurate. We expect novelty to be of influence too if built upon the accuracy of viewing interests, but the current form of our GRS is not entirely suitable to ensure novelty of the items recommended. Furthermore, diversity is a complex measurement, and interplays with accuracy and novelty on many aspects. As diversity was found to contribute to the perceived relevance of a GRS, this indicates diversity is a valued criterion for usefulness of recommendations and does not necessarily undermine a proper result. All in all, to satisfy user's needs perfectly, an interplay between several criteria seems necessary– of which accuracy, novelty, and diversity appear the most prominent actors in academic

debate (e.g. Vargas, 2011; Pu, Chen, and Hu, 2011). Further research needs to identify to what extent diversity can interplay with accuracy and novelty in order to find recommendations that are outside the box yet are still accepted by the users. Literature on interface design may help elaborate on that, as grouping of recommendations can have a considerable influence on how likely people are to recall and/or perceive them in a positive way (see: Tintarev, 2017). Similarly, Tintarev (2017) argues that spreading out recommendations that are different from the rest, instead of grouping them together, increases the perception of diversity. This could be a good starting point to elaborate on our argument.

Social relationship quality

Despite our expectations, the influence of social relationship quality (SRQ) on the watching intention of the endorsements did not yield any significant results. This could be explained by a number of issues, which will be touched upon one by one. First, we noticed a relatively high score on social relationship quality in our sample. Here we can differentiate between the two factors. The factor on SRQ similarity had a mean of 3.68 (*SD* = 0.63), on a scale ranging from 1.00 (lowest) to 5.00 (highest). The mean score for the factor on SRQ liking was even higher (*M* = 4.50, *SD* = 0.54). Perhaps not enough variance in the score on social relationship quality was found to have a significant influence on the watching intention. On the other hand, one could argue that if people have the choice, they will naturally pick a co-viewing partner that they are close with or share similar values with, hence resulting in a higher overall social relationship quality measurement.

Besides that, as established in our theoretical framework, co-viewing partners share their watching experience, and would therefore look similarly at the content provided (Tal-Or, 2019). In line with Masthoff (2011) and Hennig-Thurau, Marchand, and Marx (2012), it was argued that not all groups are alike in their level of interaction with the system and in the amount of value that is derived from a recommender system. The value that is derived from such a system could be linked to the intentions to further use the system. Following this train of thought, more value could be derived from a system if people are willing to learn from the other and reach a consensus easily, which was attributed to a higher social relationship quality as well (Jameson, 2004). As apparent from our results, it would seem as if the premise by Tal-Or (2019) could hold true for all kinds of relationships if they are above a certain social relationship quality threshold. In other words, if a certain level of social relationship quality is achieved, the dyad could be more willing to follow up on the endorsements or to make further use of the system. However, it should also be noted that the total explained variance on watching intention by social relationship quality was very low (R^2= 4.2%), meaning that there will be other measurements that are a lot stronger in predicting the watching intention of the recommended videos. Additional research is necessary to further

elaborate on these presumptions, specifically regarding further use intentions of the GRS and the boundaries of this plausible threshold. Furthermore, Tal-Or (2019) pointed out that the more liked a person is by the other, the more power he or she has in drawing somebody in and out of the storyline, which would be a supportive argument of differences resulting from variations in social relationship quality. However, since we have measured the outcome by means of watching intention, we were not able to look at actual co-viewing behaviour and potential interaction with the storyline; this would also be a potentially interesting area for future investigation.

Viewing interests of dyads

Moreover, the hypothesis on the difference or match in viewing interests of the dyads, and its impact on the watching intention also did not hold true. The mean scores on watching intention for both groups are 2.99 and 2.96 respectively, which lies almost exactly in the middle of the Likert scale, indicating a neutral perception towards watching the recommendations they received. At first sight it would seem that the GRS proved equally successful (or unsuccessful for that matter) in bridging the different perspectives of the dyads just as much as similar preferences would, however we do not want to draw this conclusion before extensively touching upon the limitations of our GRS. For starters, the results are based on a mock-up version of a GRS, which means no algorithmic structure is underlying the recommendations made. For simplicity issues, and because the average strategy is most common among dyads, it was chosen to 'build' this GRS by combining the genre preferences of the two individuals in the dyad (Masthoff, 2004). However, it is not possible to perfectly combine and average programmes based on two different genres; likely a programme will tilt slightly more towards one genre than towards the other. Since we only asked one person to fill in the questionnaire and to voice his/her opinions and perceptions on the GRS and its recommendations, skewed results are likely to emerge. That is, if the programmes that were featured in the set of recommendations tilted more towards the genre preference of the person filling out the survey, a better evaluation of the systems can be expected – and vice versa. A suggestion for further research would be to compare the scores for both individuals in a dyad, or even a group, and see how they relate to each other. Results could appear to be more of a match with one person, while being further away from the interest of the other person, but this does not necessarily have to be the case. Ideally, the recommendations would be accepted by both people, therewith giving a better indication of the effectiveness of the GRS and the suggestions it provides.

What's more, Mora, Ho, and Krider (2011) found substantial evidence to support their assumptions on varying co-viewing behaviour across programme genres in a Mexican context. Genre variations are considered to have an impact or emotional gratitude of a group of people as well. For instance, melodramatic series are generally enjoyed more when watched together, as well

as newscasts proved to be more suitable for group viewing behaviour, possible due to their mass appeal (Mora, Ho, & Krider, 2011). Other genres may prove less suitable for shared television watching situations, also dependent on genre preferences of the individuals at stake. As we investigated watching intention across six different genres, some of these could be perceived less or more suitable for group watching, therewith lowering or increasing the final score on watching intention.

Additionally, as outlined by Mora, Ho, and Krider (2011), conflict management may be necessary in case of opposing viewing interests, which was not addressed by our survey. Since we have asked participants to fill out the choice of genre for the other person, they are very conscious about a mismatch in those preferences, if there was any. Tal-Or (2016) pointed out that people tend be very aware of their shared media consumption as they conversate about their media intake. It would be plausible that because the participant can anticipate this mismatch, it (subconsciously) prepares him/her for being exposed to different viewing suggestion than they would usually expect, which could in turn lead to a more accepting attitude towards the endorsements. A limited account of conflict management could also be a consequence of the natural tendency to learn about each other, and the interests of the other, as was outlined by both Jameson (2004) and Tal-Or (2019). That is, in case of distinct viewing preferences, people that know and appreciate the other person well for who they are and what they are interested in, may be more appreciative of the combined viewing suggestions, as they are based on both of their inputs. Here, we can draw a connection with the social relationship quality aspect, while explaining the little variance among the scores for watching intention of both groups. Besides that, similar to our argument for social relationship quality, one could argue that if people have the choice, they are likely to pick a co-viewing partner with a similar psychographic viewer profile. According to Lull (1980)'s theory on Social Uses of TV, corresponding profiles would foster co-viewing intentions, despite potential differences in genre preferences. Here, the closeness to one's viewing partner outweighs the chosen content to watch, when it comes to the co-viewing experience. Lastly, as a methodological limitation, one could also argue that since the participant was encouraged to continue the questionnaire in case of non-response based on what they *think* the other person would want to watch, the results could be slightly more favourable to the viewing preference of the respondent.

5.2 Limitations and suggestions for further research

In addition to the limitations mentioned previously, some methodological implications will be discussed below, as well as suggestions to overcome those in future studies. To start with, the most prominent limitation of this research is that it is based on a mock-up version of a GRS, not on

reality. Of course, no such group recommender system is in place, especially with regards to public broadcasting or the NPO in specific. Albeit a relatively unavoidable limitation for this amount of time and resources, it does mean some remarks should be placed alongside our outcomes. Although the suggestions were based on existing television content, it was a random selection made from the perspective of the researcher, which in any case will be biased (Malhotra, 2006). The small-sampled test before conducting the actual research helped to ensure that the programme-genre fit was accurate, yet the sample was likely to be in the same socioeconomic sphere as the researcher. The current results are based on a very specific selection of TV shows, which also means our (non-) significant findings may not hold true for all (combinations of) recommendations. That is, replacing one or more of the TV shows with another may lead to substantially different outcomes. Replications of this study with other input will have to show whether or not this is the case. Furthermore, no previous watching history or viewing preferences, liking/disliking of content whatsoever could be taken into account when providing the participants with their recommendations, hence portraying a less accurate system than any group recommender system's algorithm would be able to produce. Yet, mimicking an algorithm was never the intention of the study, but it would be good starting point for a follow up study, to see if the algorithm is able to generate the same results. With that, as an algorithm is based upon actual user activity, it should provide recommendations that form a better connection with the interests of its users. Thus, as one would expect these suggestions to be more accurate, and to a certain extent novel, than those provided by our GRS, it would be interesting to see whether users still indicate similar ratings for watching intention or further use intention of the algorithm-based GRS, or if they increase. Besides that, an algorithm may also be better at finding the right balance between accuracy, novelty, and diversity; this could then greatly affect the mediating power of perceived usefulness on the relationship between the intentions for further usage and intentions to watch the endorsed videos. Qualitative or mixed methods could also be used to identify whether more criteria are essential to the user's evaluation of the usefulness of the recommendations by such a GRS.

Secondly, some issues with the validity and reliability of our variables have arisen. Two of our constructed scales – social relationship quality and perceived usefulness – were found to have a (very) low Cronbach's alpha score, which is the most common indicator of reliability in behavioural survey research (Shelby, 2011). Although Shelby (2011) has put forward some critique on using this measurement to assess reliability and the scales did pass his suggested alternative criterium on a correction item-total correlation smaller than .40 – we cannot neglect the relatively weak correlation of the items in the scale. Furthermore, Shelby (2011) pointed out that when measuring complex human dimensions, such as attitudes, perceptions or emotions, it is often forgotten that a high variation among groups of individuals in a population is expected. This will again be reflected

in the scores of a scale, which can also explain the difference in outcomes of reliability analyses between our sample, and those of the previous uses of the same (valid) scale. Another explanation for a lower Cronbach's alpha score could be derived from the fact that the scales were translated into another language, which may have unintentionally affected the connotations of certain phrasing. Moreover, the construct of relevance was measured through further use intentions since there was no room in our current design for choosing whether or not to make use of the system – one simply had to imagine they were. Hence, it was decided to ask for intentions to further use the system to indicate whether users thought of it as a relevant tool that they would be willing to adopt in the future. However, this measurement could be influenced by the participant's perceptions of this mock-up GRS, rather than any GRS in general. This could have had both a negative and a positive effect on the score for further use intentions. Furthermore, the participants' perception on relevance of a GRS may be subject to other external influences, such as the setting they are watching, mood etcetera. In this light, relevance and further use intentions may vary slightly, so we suggest that the former will be investigated more extensively in follow-up research. Nonetheless, the construct on further use intentions did allow us to draw some interesting conclusions. Future studies are encouraged to look at the construct perceived relevance more critically, and perhaps find a way to measure this at the beginning of the user's experience with a (mock-up) GRS. A suggestion would be to create an experiment, as this will prove useful in creating groups to draw comparisons between. This study already borrows some elements from experimental designs by comparing dyads with similar or different genre preferences, and by the inclusion of a manipulation check. It was decided to continue this research using cross-sectional survey research, since this was considered to be a less complex approach and better suitable method for testing the hypotheses that were brought forth. Masthoff (2011) already established the benefits of conducting experiments in assessing the effectiveness of (group) recommender systems, thus we propose to give people a choice in deciding whether or not to make use of the GRS, in order to assess the user intentions at the exact moment of engaging in co-viewing behaviour and deciding on what programme to watch together. Besides the perceived relevance of the system, this would also make room to address the visual attractiveness or user-friendliness of the design.

Thirdly, some limitations regarding the generalizability of our study should be highlighted. Although our sample was big enough to be able to conduct proper statistical analyses, our target population was quite broad, making it easier for skewed demographics to arise. In our study, the sample was relatively young and high educated, which among Dutch society only accounts for a few percent of the people. The question can thus be raised how well our sample can be representative of the target audience of the NPO. On the other hand, based on studies on why people make use of online video streaming in the digital age, it was found that youngsters and most people in their

thirties have moved away completely from cable television, and access all their television content online (Lagger, Lux, & Marques, 2017). Based on this knowledge, our target group may prove to be a reasonable depiction of the age distribution among contemporary users of online streaming platforms, as they generally attract younger audiences.

5.3 Strengths

Lastly, this research also carries considerably strengths that should be pointed out, as they add to the existing debate on the social implications of group recommender systems and co-viewing behaviour. First of all, this research is based on an elaborate theoretical basis, not only guiding the literary framework for this study, but also most of the methodological decisions that have been made, which make this research suitable for repeated measurements.

Furthermore, this research was supported by the NPO, which opened access to many resources that improved the reliability of our measurements. This could be considered especially important because of the hypothetical nature of our study. The GRS system as we have investigated is not currently in use by any (public) broadcaster in the Dutch media landscape, which makes this research highly contemporary and relevant, but also harder to generalize. Therefore, this research clearly benefited from the widely known and respectable name of the NPO among Dutch society. Also, the use of real, and for a large part familiar television programmes, helped in creating a setting that looked familiar and more tangible. Similarly, it was highly beneficial for the reliability of the proposed GRS system to be able to show an interface that is duplicating the current style of the NPO.

5.4 Practical and scientific implications

As this research was written with the help of the NPO and their resources, our study has provided some practical and scientific implications that could benefit the development of algorithms for group recommendations and the underlying criteria for endorsement.

By means of this study, we have established the mediating factor of the perceived usefulness criteria of accuracy on the relationship between further use intentions of our mock-up GRS and the watching intention of endorsed online public broadcasting content. This result was not significantly different among the match or mismatch in viewing interests of the dyads nor was it influenced by the social relationship quality of the dyads. Although no evidence was found for the influence of novelty and diversity – as constructs for the perceived usefulness of the recommendations – on the watching intention, we discussed how this could have been explained theoretically and/or methodologically. Yet, these results also carry a vast amount of practical relevance, as well as it ties in with the aforementioned paternalism – popularity debate (Sørensen

& Hutchinson, 2017).

Public service broadcasting is known for its societal purpose to educate in its widest form and foster open-mindedness of the public (Sørensen & Hutchinson, 2017). These broadcasters are supposed to attract niches and the general audience with content that can be perceived out-of-the-box and diverse, resulting from their main function in educating their audiences. However, as is expressed through the paternalism – popularity debate, the recommendations that public broadcasters are providing are ought to be highly accurate in relation to user's previous watching behaviour; if not, the lack of transparency on the rationale behind the recommendation can easily turn into an accusation of recommending for their own sake instead of for that of the user. Here the interplay between accuracy and diversity seems most apparent.

This debate has prolonged in the development of algorithmic structures, specially designed for public broadcasting. As was derived earlier, accuracy on its own is not enough to build an algorithm because of the questionable relevance of the isolated criterium. As became apparent through this research, diversity did not prove to be a good predictor for watching intention of the recommenders by a GRS, and thus the effectiveness of the system created. What's more, no automated recommender system is able to portray and understand the values of diversity in a way traditional broadcasting was able to do through programming, production and distribution (Sørensen, & Hutchinson, 2017). This has serious consequences for creating group recommender systems since algorithms are based on mathematical calculations that strive for optimisation of personalised recommendations. A key issue in the public service debate is one at the crossing between the collective social service function of public broadcasting and tailored content for individual consumers (as would be the case for commercial broadcasters; Sørensen, & Hutchinson, 2017). Perhaps the introduction of group recommender systems for public service media will make it possible to generate better contributions to the societal collective, thus easing the latter tension.

Nevertheless, Tintarev (2017) has developed a diversity-aware recommendation model, considering both item and user diversity, which can be deducted from previous user activity and ratings. The prevailing question guiding this model is how to minimize potential polarization of opinions, while ensuring trust and providing the most diverse outcome as possible. Perhaps, a logical next step would be to apply this diversity aware recommendation model to a group setting, and to the specific case of the Dutch public broadcaster NPO. A beneficial factor for public service recommendation tactics could be found in the limitations of online (search) engine recommendations. These recommendations are based on textual descriptions (metadata) which do not entirely match the semantic content of, for instance, a television programme (Zhou, Khemmarat, Gao & Wang, 2011). In many cases, textual descriptions are consistent of titles and tags only, which restricts the extent to which the full magnitude of a show can be exposed.

However, textual metadata is hence easier to manipulate in order to boost popularity of one item over another, which could be of great use to match preferences. With regards to public broadcasting companies that are in charge of producing their very own content, this might come with additional benefits. Since they are ought to teach the public on a wider range of perspectives and foster open-mindedness to the unknown (Sørensen, & Hutchinson, 2017), public broadcasters may want to make use of this 'flaw' for their own good. Co-viewers with very distinct preferences, but who have a close social relationship to each other, may be open to learn from the other, as pointed out by Jameson (2004) before; a standpoint that could be used as a reference point for providing them with a broader range of viewing suggestions in the future.

Thus, by means of this study we argue that inserting diversity within group recommender systems is a challenging task, but one that has enough benefit to the non-commercial intentions of public service broadcasting that it is worth investing in. By letting the commercialised approach to recommending thrive and providing the audience with reinstatements of their own interests and beliefs by building upon an accuracy-led approach, one is putting the open-minded and well-read citizen and the future of public service media at risk. If accuracy takes the upper hand in recommender system development, it will become harder for public broadcasters to voice their opinions, and share their valuable content, since people are likely to stick to their own filter bubble. We realize that a fully equipped diversity-aware (group) recommendation model is not quite there yet, nor is the proper model for assessing the effectiveness of such a system (e.g. in terms of watching intention or perceived relevance), but we hope this work has contributed to establishing a groundwork for the importance of continuing research in this field.

References

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2013). Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Information Systems Research, 24*(4), 956-975. http://dx.doi.org/10.1287/isre.2013.0497

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502-508. http://dx.doi.org/10.1111/opo.12131

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology, 63*(4), 596-612. https://doi.org/10.1037/0022-3514.63.4.596

Baltar, F., & Brunet, I. (2012). Social research 2.0: virtual snowball sampling method using Facebook. *Internet research, 22*(1), 57-74. https://doi.org/10.1108/10662241211199960

Bardoel, J. (2003). Back to the public? Assessing public broadcasting in the Netherlands. *Javnost-The Public, 10*(3), 81-96. https://doi.org/10.1080/13183222.2003.11008836

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research – Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. Retrieved from http://webcom.upmf-grenoble.fr/LIP/Perso/DMuller/GSERM/Articles/Journal%20of%20Personality%20and%20Social%20Psychology%201986%20Baron.pdf

Bellman, S., Robinson, J. A., Wooley, B., & Varan, D. (2017). The effects of social TV on television advertising effectiveness. *Journal of Marketing Communications, 23*(1), 73-91. https://doi.org/10.1080/13527266.2014.921637

Bennett, J., & Lanning, S. (2007, August). The Netflix Prize. *Proceedings of KDD Cup and Workshop 2007*. Retrieved from https://www.cs.uic.edu/~liub/KDD-cup-2007/NetflixPrize-description.pdf

Blumler, J. G., & Katz, E. (1974). Utilization of the mass communication by the individual. In J. G. Blumler & E. Katz (Eds.), *The uses of mass communications: Current perspectives on gratifications research.* Beverly Hills, CA: Sage Publications.

Bollen, D., Knijnenburg, B. P., Willemsen, M. C., & Graus, M. (2010, September). Understanding choice overload in recommender systems. *Proceedings of the fourth ACM conference on Recommender systems*, 63-70. https://doi.org/10.1145/1864708.1864724

Chaney, A. J., Gartrell, M., Hofman, J. M., Guiver, J., Koenigstein, N., Kohli, P., & Paquet, U. (2014, June). A large-scale exploration of group viewing patterns. *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, 31-38. https://doi.org/10.1145/2602299.2602309

Chorianopoulos, K. (2007). Content-enriched communication-supporting the social uses of TV. *Journal of The Communications Network, 6(*1), 23-30. Retrieved from https://pdf.epidro.me/Chorianopoulos_2007e.pdf

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T. (2010, September). The YouTube video recommendation system. *Proceedings of the fourth ACM conference on Recommender systems*, 293-296. https://doi.org/10.1145/1864708.1864770

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction, 4*(2), 81-173. https://doi.org/10.1561/1100000009

Geerts, D., & De Grooff, D. (2009, April). Supporting the social uses of television: sociability heuristics for social TV. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 595-604. https://doi.org/10.1145/1518701.1518793

Girgensohn, A. and Lee, A. (2002, November). Making web sites be places for social interaction. *Proceedings of the 2002 ACM conference on Computer supported cooperative work,* 136-145. https://doi.org/10.1145/587078.587098

Gomez-Uribe, C. A., & Hunt, N. (2015). The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems, 6*(4), 1-19. https://doi.org/10.1145/2843948

Haridakis, P., & Hanson, G. (2009). Social interaction and co-viewing with YouTube: Blending mass communication reception and social connection. *Journal of Broadcasting & Electronic Media*, *53*(2), 317-335. https://doi.org/10.1080/08838150902908270

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach* (2nd edition). New York : The Guilford Press

Hennig-Thurau, T., Marchand, A., & Marx, P. (2012). Can automated group recommender systems help consumers make better choices? *Journal of Marketing*, *76*(5), 89-109. https://doi.org/10.1509/jm.10.0537

Horenberg, D. (2019). To what extent does the use of a group recommender system positively affect a dyad's intention to watch an endorsed video? A study examining the effectiveness of group recommender systems. [Master's thesis]. Retrieved from http://www.ubvu.vu.nl/pub/fulltext/scripties/27_2652921_0.pdf

Jameson, A. (2004, May). More than the sum of its members: Challenges for group recommender systems. *Proceedings of the working conference on Advanced visual interfaces*, 48-54. https://doi.org/10.1145/989863.989869

Janssen, S., & Verboord, M. (2017). Methodological guidelines thesis research, version 4. Department of Media and Communication, Erasmus School of History, Culture and

Communication. Retrieved from https://www.eshcc.eur.nl/english/media

Knijnenburg, B. P., Willemsen, M. C., & Hirtbach, S. (2010). Receiving recommendations and providing feedback: The user-experience of a recommender system. In F. Buccafurri & G. Semeraro (Eds.), *E-Commerce and Web Technologies* (pp. 207-216). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-15208-5_19

Köcher, S., Jugovac, M., Jannach, D., & Holzmüller, H. H. (2019). New hidden persuaders: An investigation of attribute-level anchoring effects of product recommendations. *Journal of Retailing, 95*(1), 24-41. https://doi.org/10.1016/j.jretai.2018.10.004

Lagger, C., Lux, M., & Marques, O. (2017). What makes people watch online videos: An exploratory study. *Computers in Entertainment (CIE), 15*(2), 1-31. http://dx.doi.org/10.1145/3034706

Lee, M. S., Heeter, C., & LaRose, R. (2010). A modern Cinderella story: a comparison of viewer responses to interactive vs linear narrative in solitary and co-viewing settings. *New Media & Society*, *12*(5), 779-795. Doi: 10.1177/1461444809348771

Lee, B. and Lee, R. S. (1995). How and why people watch TV: Implications for the future of interactive television. *Journal of Advertising Research, 35*(6), 9-18.

Lull, J. (1980). The social uses of television. *Human Communication Research, 6*(3), 197-209. https://doi.org/10.1111/j.1468-2958.1980.tb00140.x

Malhotra, N. K. (2006). Questionnaire design and scale development. In R. Grover & M. Vriens (Eds.), *The handbook of marketing research: Uses, misuses, and future advances* (pp. 83-94). Thousand Oaks: SAGE. Retrieved from https://www.researchgate.net/profile/Naresh_Malhotra/publication/266864633_Questionnaire_design_and_scale_development/links/566708fe08ae34c89a0220f9.pdf

Masthoff, J. (2004). Group modeling: Selecting a sequence of television items to suit a group of viewers. In *Personalized digital television* (pp. 93-141). Dordrecht: Springer. https://doi.org/10.1007/1-4020-2164-X_5

Masthoff, J. (2011). Group recommender systems: Combining individual models. In F. Ricci et al. (Eds.), *Recommender systems handbook* (pp. 677-702). Boston: Springer. Doi: 10.1007/978-0-387-85820-3_21

Matthews, B. & Ross, L. (2010). C3: Questionnaires. In B. Matthews & L. Ross (Eds.), *Research Methods: A practical guide for the social sciences* (pp. 200-217). Harlow: Pearson.

Mora, J. D., Ho, J., & Krider, R. (2011). Television co-viewing in Mexico: An assessment on people meter data. *Journal of Broadcasting & Electronic Media, 55*(4), 448-469. https://doi.org/10.1080/08838151.2011.620905

Nagulendra, S., & Vassileva, J. (2014, September). Understanding and controlling the filter bubble through interactive visualization: a user study. *Proceedings of the 25th ACM conference on Hypertext and social media*, 107-115. http://dx.doi.org/10.1145/2631775.2631811

Neuman, W.L. (2014). Experimental Research. In W.L. Neuman (Ed.), *Social Research Methods: Qualitative and Quantitative Approaches* (7th edition) (pp. 281-313). Essex: Pearson.

NPO. (n.d.). Our mission. [NPO mission statement]. Retrieved from https://over.npo.nl/organisatie/about-npo/our-mission#content

Paavonen, E. J., Roine, M., Pennonen, M., & Lahikainen, A. R. (2009). Do parental co-viewing and discussions mitigate TV-induced fears in young children? *Child: care, health and development, 35*(6), 773-780. http://dx.doi.org/10.1111/j.1365-2214.2009.01009.x

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications, 97*, 205-227. http://dx.doi.org/10.1016/j.eswa.2017.12.020

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. *Proceedings of the fifth ACM conference on Recommender systems*, 157-164. http://dx.doi.org/10.1145/2043932.2043962

Quijano-Sánchez, L., Díaz-Agudo, B., & Recio-García, J. A. (2014). Development of a group recommender application in a social network. *Knowledge-Based Systems*, *71*, 72-85. http://dx.doi.org/10.1016/j.knosys.2014.05.013

Rubin, A. M. (1983). Television uses and gratifications: The interactions of viewing patterns and motivations. *Journal of Broadcasting, 27*(1), 37-51. http://dx.doi.org/10.1080/08838158309386471

Shelby, L. B. (2011). Beyond Cronbach's alpha: Considering confirmatory factor analysis and segmentation. *Human dimensions of wildlife, 16*(2), 142-148. http://dx.doi.org/10.1080/10871209.2011.537302

Silveira, T., Zhang, M., Lin, X., Liu, Y., & Ma, S. (2019). How good your recommender system is? A survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics, 10*(5), 813-831. http://dx.doi.org/10.1007/s13042-017-0762-9

Skouteris, H., & Kelly, L. (2006). Repeated-viewing and co-viewing of an animated video: an examination of factors that impact on young children's comprehension of video content. *Australian Journal of Early Childhood, 31*(3), 22-30. http://dx.doi.org/10.1177/183693910603100305

Sørensen, J. K., & Hutchinson, J. (2017). Algorithms and public service media. In G.F. Lowe, H. Van den Bulck, & K. Donders (Eds.), *Public Service Media in the Networked Society RIPE@2017*, (pp. 91-106). http://dx.doi.org/10.5167/uzh-159817

Strouse, G. A., Troseth, G. L., O'Doherty, K. D., & Saylor, M. M. (2018). Co-viewing supports toddlers' word learning from contingent and noncontingent video. *Journal of experimental child psychology, 166*, 310-326. http://dx.doi.org/10.1016/j.jecp.2017.09.005

Tal-Or, N. (2016). How co-viewing affects attitudes: The mediating roles of transportation and identification. *Media Psychology*, *19*(3), 381-405. http://dx.doi.org/10.1080/15213269.2015.1082918

Tal-Or, N. (2019). The Effects of Co-Viewers on the Viewing Experience. *Communication Theory, 00*, 1-20. http://dx.doi.org/10.1093/ccc/qtz012

Tintarev, N. (2017, August 31). Presenting diversity aware recommendations: Making challenging news acceptable. *Proceedings of The FATREC Workshop on Responsible Recommendation*. http://dx.doi.org/10.18122/B2HQ41

Trattner, C., Said, A., Boratto, L., & Felfernig, A. (2018). Evaluating group recommender systems. In A. Felfernig, L. Boratto, M. Stettinger, & M. Tkalčič (Eds.), *Group Recommender Systems: An Introduction* (pp. 59-71). Cham: Springer. https://doi.org/10.1007/978-3-319-75067-5_3

Tryon, C. (2015). TV got better: Netflix's original programming strategies and binge viewing. *Media Industries Journal, 2*(2), 104-116. Retrieved from https://quod.lib.umich.edu/cgi/p/pod/dod-idx/tv-got-better-netflixs-original-programming-strategies.pdf?c=mij;idno=15031809.0002.206;format=pdf

Vargas, S. (2011, August). New approaches to diversity and novelty in recommender systems. *Fourth BCS-IRSG Symposium on Future Directions in Information Access (FDIA 2011), 4*, 8-13. Retrieved from https://www.scienceopen.com/document_file/4a653130-e70f-4bbe-9fe4-fcd0a5aeefb2/ScienceOpen/008_Vargas.pdf

Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478. Retrieved from https://www.jstor.org/stable/pdf/30036540.pdf?casa_token=2un82eH6MwoAAAAA:bMuoK-dezW0P6jPfQ_Gtvu88yhZmfVKeYFTUvFWsxqHW6ueFlRT6XCW5EXWQruyDB1_6PTqg8lxSQs3BF4tEYUqml6Iief0S1jttjxRbnkghoIiwPYUcAA

Xue, Y., Moitra, A., & Gustafson, S. M. (2013). Methods and systems for online recommendation. U.S. Patent No. 8,365,227. Washington, DC: U.S. Patent and Trademark Office. Retrieved from https://patentimages.storage.googleapis.com/9f/1e/34/e03139d8a667e8/US8365227.pdf

Young, M. (2020, February 14). Does distance make the heart grow fonder? [Blog post]. Retrieved from https://www.sexlab.ca/blog?offset=1589469692982

Zajonc, R. B. (1965). Social facilitation. *Science, 149*(3681), 269-274.

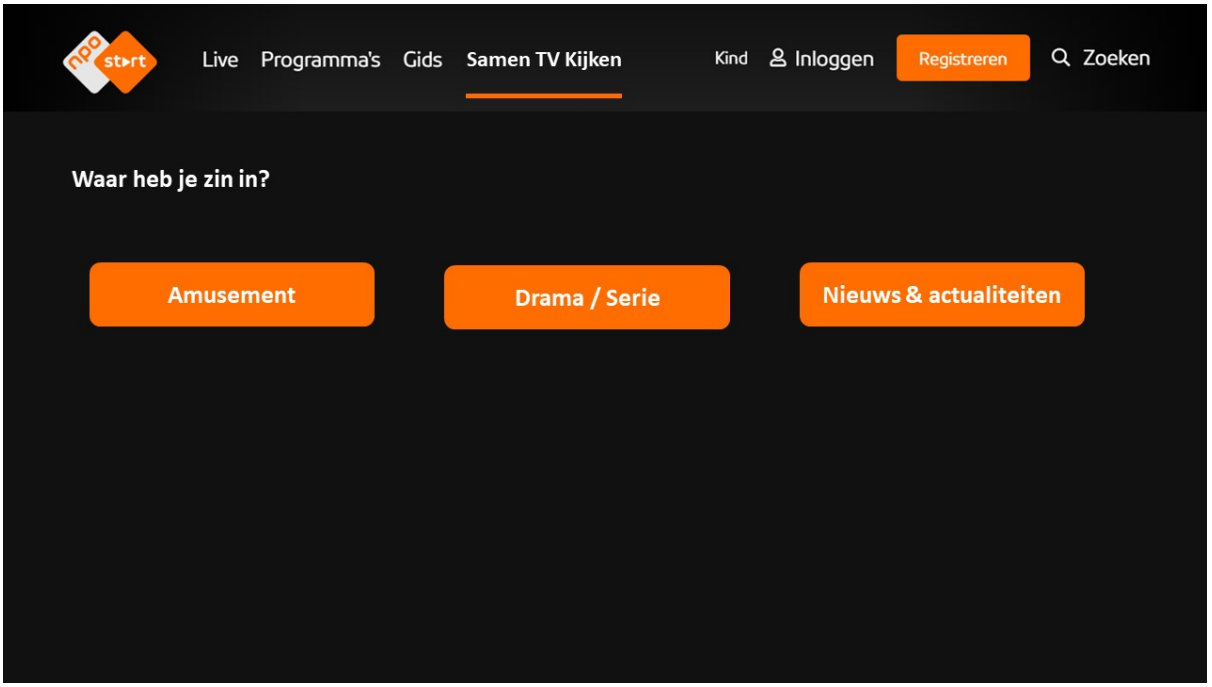
http://dx.doi.org/10.1126/science.149.3681.269

Zhang, J. (2011, October). Anchoring effects of recommender systems. *Proceedings of the fifth ACM conference on Recommender systems,* 375-378. http://dx.doi.org/ 10.1145/2043932.2044010

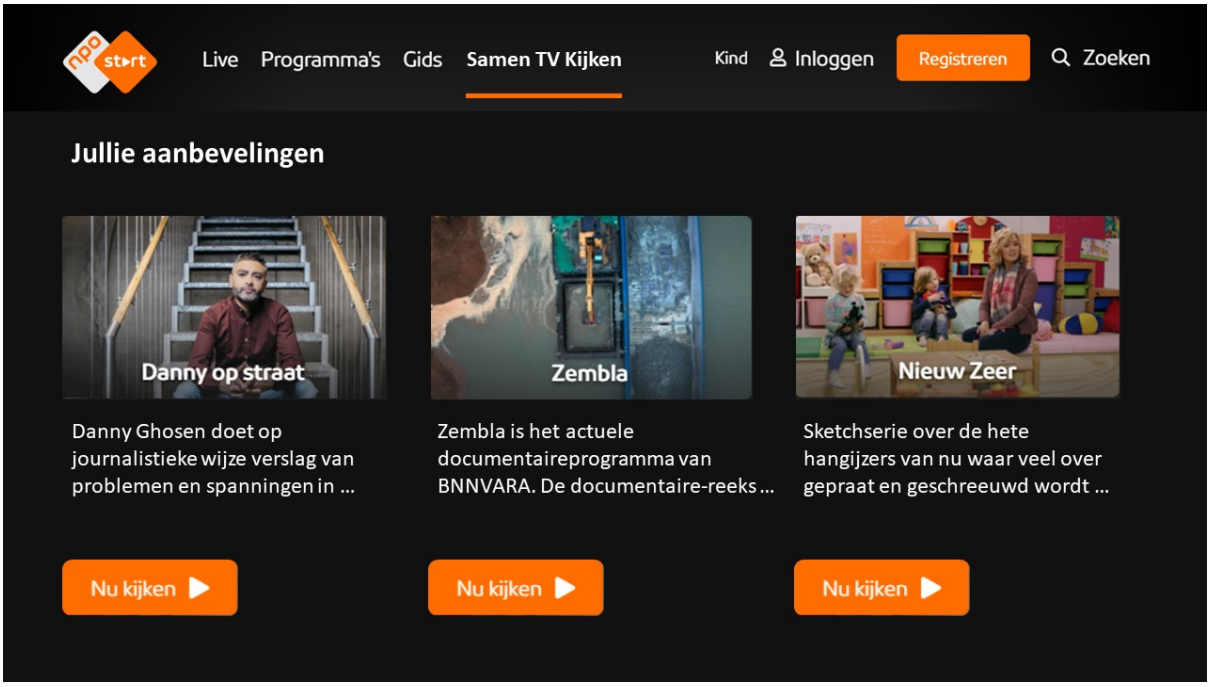
Zhang, L. (2013). The definition of novelty in recommendation system. *Journal of Engineering Science and Technology Review, 6*(3), 141-145. Retrieved from http://www.jestr.org/downloads/Volume6Issue3/fulltext25632013.pdf?utm_campaign=elearningindustry.com&utm_source=%2Ffuture-of-learning-tired-waiting-click-now&utm_medium=link

Zhou, R., Khemmarat, S., Gao, L., & Wang, H. (2011). Boosting video popularity through recommendation systems. *Databases and Social Networks*, 13-18. http://dx.doi.org/10.1145/1996413.1996416

Appendix A: Interface designs NPO Start

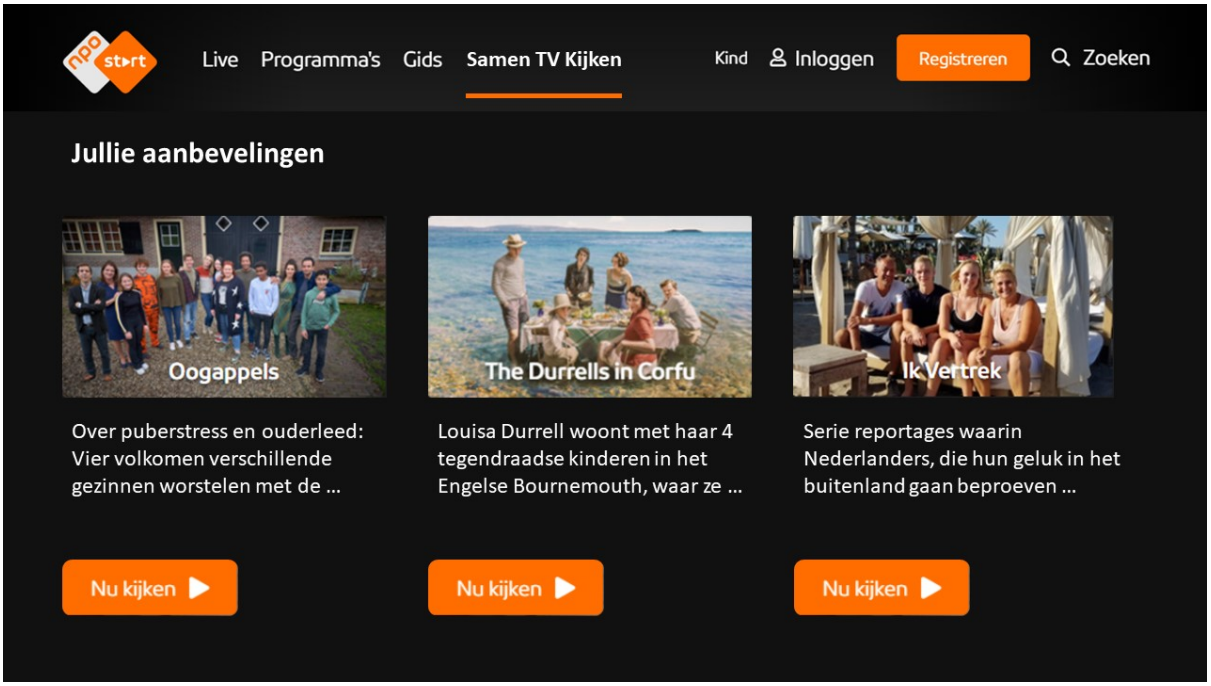


Appendix A1: Interface design for the question "Welk genre zou jij/de ander het liefst willen kijken?" [Which genre do you/the other person want to watch?]

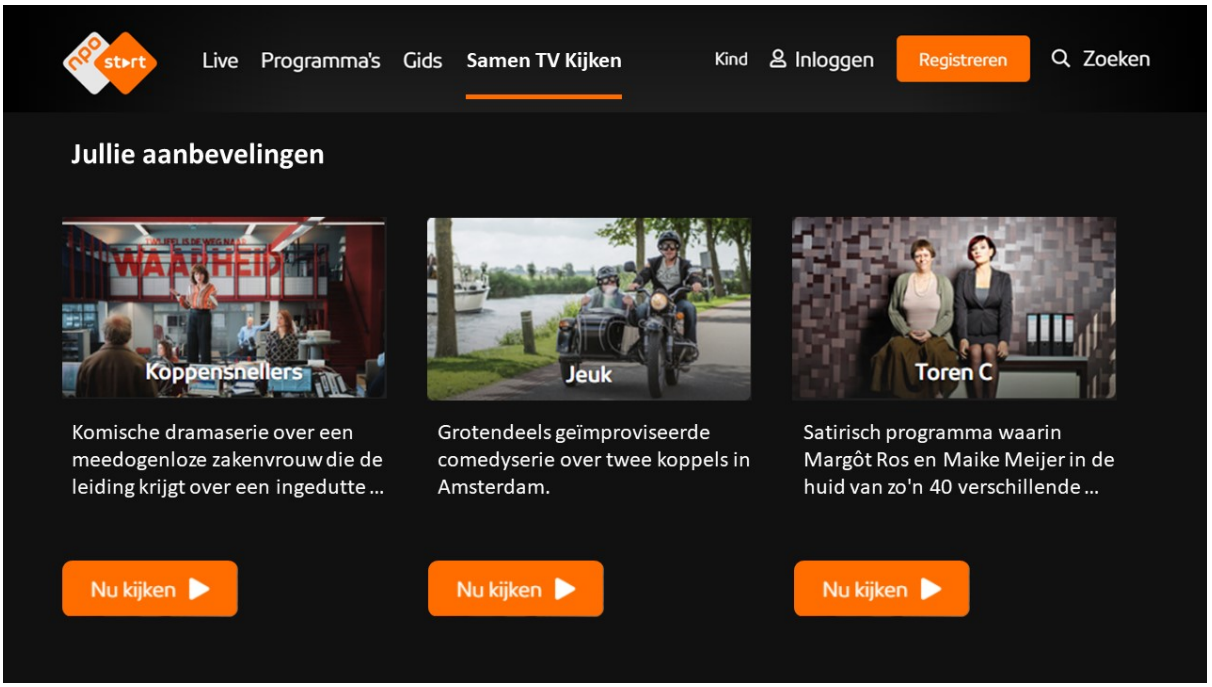


Appendix A2: Interface design mock-up GRS' selection of recommended videos for Topical interest

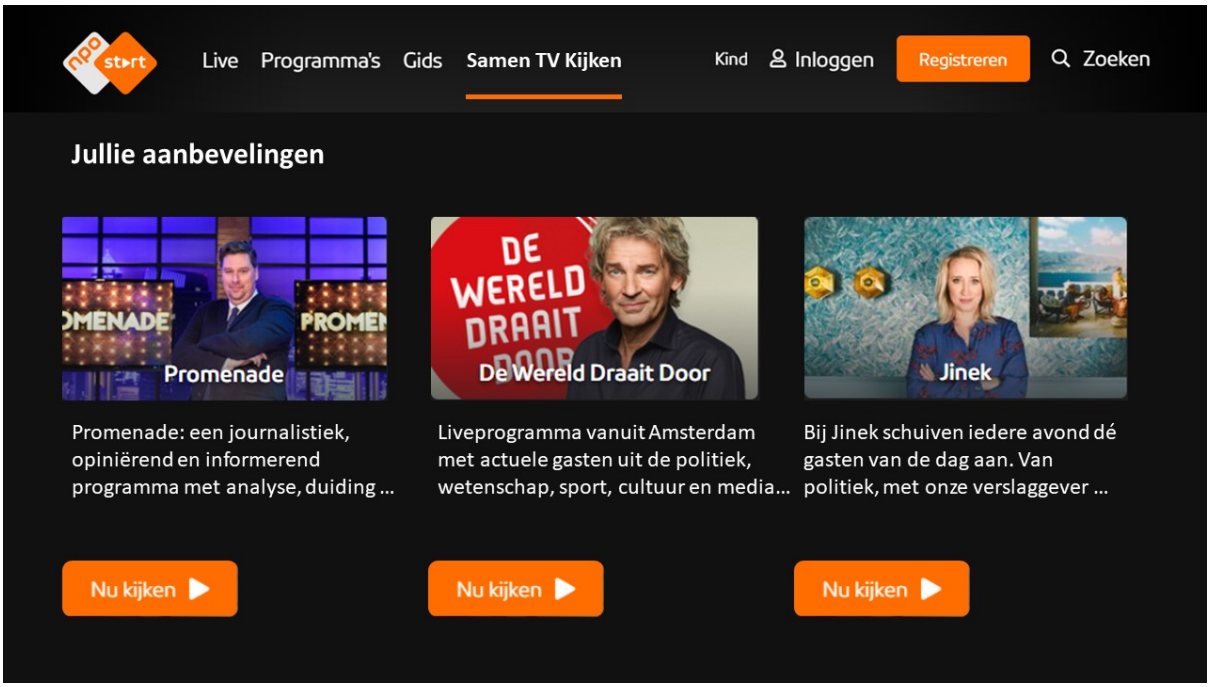
Appendix A3: Interface design mock-up GRS' selection of recommended videos for Entertainment



Appendix A4: Interface design mock-up GRS' selection of recommended videos for Drama series



Appendix A5: Interface design mock-up GRS' selection of recommended videos for Humorous drama



Appendix A6: Interface design mock-up GRS' selection of recommended videos for News & current events

Appendix A7: Interface design mock-up GRS' selection of recommended videos for Satire

Appendix B: Qualtrics survey questions (in Dutch)

Beste participant,

Let op: probeer deze enquête in te vullen met de voorkeuren van iemand waarmee je regelmatig samen televisie kijkt

Hartelijk dank voor het deelnemen aan deze enquête, die bijdraagt aan een master scriptie van de Master in Media and Business aan de Erasmus Universiteit Rotterdam. Door middel van dit onderzoek willen we graag betere inzichten krijgen in de hulpmiddelen die je zouden kunnen helpen bij het gezamenlijke keuzeproces van een televisieprogramma waarin beide kijkers interesse hebben. Vandaar dat we graag een realistische situatie schetsen en de voorkeuren willen meenemen van iemand waarmee je regelmatig samen televisie kijkt (18 jaar of ouder; bijvoorbeeld partner, huisgenoot, familie, vriend). Je wordt aangemoedigd om iemand hiervoor daadwerkelijk te benaderen, dit mag zowel fysiek in dezelfde ruimte of digitaal/telefonisch, maar mocht dit niet mogelijk zijn, kun je de vragenlijst ook hervatten met de informatie gebaseerd op wat je *denkt* dat de ander zou willen kijken.

Je participatie in dit onderzoek is geheel vrijwillig; je kunt dus op ieder moment de enquête stopzetten of weer hervatten. Ook heb je het recht om sommige vragen onbeantwoord te laten. Voor zover bekend zitten er geen risico's aan je deelname: je persoonlijke informatie wordt strikt vertrouwelijk behandeld en uitsluitend gebruikt voor academisch onderzoek, waarbij anonimiteit volledig wordt gegarandeerd. Het voltooien van de vragenlijst duurt ongeveer 5-10 minuten. Mochten er vragen zijn tijdens of na je participatie, voel je vrij om Jessica Broeders te contacteren (432482jb@student.eur.nl).

O Ik heb bovenstaande informatie begrepen, en ga akkoord met mijn deelname aan dit onderzoek.

Voordat we beginnen, willen we graag een aantal dingen weten over je achtergrond en persoonlijk gebruik van online streamingdiensten.

Wat is je leeftijd?

▼ Jonger dan 18 ... Ouder dan 70

Welk geslacht identificeer je je mee?

- O Man
- O Vrouw
- O Anders
- O Wil ik liever niet zeggen

Wat is je hoogst genoten opleiding?

- O Voortgezet onderwijs (VMBO, MULO, MAVO, HAVO, VWO, etc.)
- O Lager beroepsonderwijs (LTS, LEAO, LHNO, etc.)
- O Middelbaar beroepsonderwijs (MBO)
- O Hoger beroepsonderwijs (HBO)
- O Universitair onderwijs (WO Bachelor)
- O Hoger universitair onderwijs (WO Master)
- O Doctoraat, MBA, of vergelijkbaar
- O Anders, namelijk: _____

Helaas kun je deze enquete alleen invullen als je 18 jaar of ouder bent. Toch bedankt voor je interesse!

Voor dit onderzoek wordt een interface gebruikt vergelijkbaar met die van NPO Start. Voor degenen die niet bekend zijn met deze applicatie, NPO Start is een online streamingdienst waar men films of afleveringen van series terug kan kijken, die eerder verschenen zijn op één van de drie NPO televisie kanalen (NPO 1, 2 en 3).

Welke online streamingdiensten heb je wel eens gebruikt?
(Meerdere antwoorden mogelijk)

- ☐ NPO Start
- ☐ Netflix
- ☐ HBO
- ☐ Disney+
- ☐ Videoland
- ☐ Amazon Prime Video
- ☐ NL Ziet
- ☐ Ziggo Movies & Series XL
- ☐ Film1
- ☐ Apple TV+
- ☐ Anders, namelijk: _____

Ook zouden we graag wat te weten komen over iemand waarmee je frequent samen televisie kijkt. *[Als je dit nog niet had gedaan, is dit een goed moment om, fysiek of digitaal, contact te zoeken met iemand].*

Met wie ga je deze enquête invullen?

- O Partner (gehuwd, geregistreerd partnerschap, samenlevingscontract etc.)
- O Familielid
- O Vriend/vriendin (vriendschappelijke relatie)
- O Huisgenoot
- O Collega
- O Anders, namelijk: _____

In bovenstaande afbeelding zie je telkens twee cirkels die naast elkaar liggen of elkaar overlappen, met daarin steeds een verwijzing naar jezelf (Self) en de ander (Other); in dit geval degene waarmee je deze enquête gaat invullen. Hoe verder de cirkels elkaar overlappen representeert een innige of diepgaande relatie tot de ander. Hoe verder de cirkels uit elkaar liggen representeert een meer oppervlakkige relatie.

Gebaseerd op de combinatie van cirkels in bovenstaande afbeelding, hoe zou je je relatie tot de ander beschrijven?

- O Zoals combinatie 1
- O Zoals combinatie 2
- O Zoals combinatie 3
- O Zoals combinatie 4
- O Zoals combinatie 5
- O Zoals combinatie 6
- O Zoals combinatie 7

Geef aan in hoeverre je het eens bent met de volgende stellingen.

	Helemaal oneens	Oneens	Neutraal	Eens	Helemaal eens
Ik waardeer mijn tv-partner erg als persoon	O	O	O	O	O
Ik zie mijn tv-partner als een goede vriend	O	O	O	O	O
Ik kan het goed vinden met mijn tv-partner	O	O	O	O	O

Geef aan in hoeverre je het eens bent met de volgende stellingen.

	Helemaal oneens	Oneens	Neutraal	Eens	Helemaal eens
Mijn tv-partner en ik zijn op veel gebieden hetzelfde	O	O	O	O	O
Mijn tv-partner en ik kijken op een vergelijkbare manier naar veel dingen	O	O	O	O	O
Mijn tv-partner en ik zijn overeenkomend als het gaat om onze zienswijze, opvattingen en waarden	O	O	O	O	O

Stel je voor dat je nu televisie zou gaan kijken met de persoon waarmee je contact hebt gezocht. De interface van NPO Start zou er dan zo uit kunnen zien als hierboven.

Welk genre zou **jij** het liefst willen kijken?

- O Amusement
- O Drama / Serie
- O Nieuws & actualiteiten

Welk genre zou **de ander** het liefst willen kijken?
Of, als het niet mogelijk was om iemand anders te bereiken, welk genre *denk* je dat de ander het liefst zou willen kijken?

- O Amusement
- O Drama / Serie
- O Nieuws & actualiteiten

Als laatste onderdeel waarbij de bijdrage van de ander nodig is, zouden we graag dezelfde drie demografische vragen stellen aan degene waarmee je frequent samen televisie kijkt.

Wat is de leeftijd van de ander?

▼ Jonger dan 18 ... Ouder dan 70

Welk geslacht identificeert de ander zich mee?

- O Man
- O Vrouw
- O Anders
- O Wil ik liever niet zeggen

Wat is de hoogst genoten opleiding van de ander?

- O Voortgezet onderwijs (VMBO, MULO, MAVO, HAVO, VWO, etc.)
- O Lager beroepsonderwijs (LTS, LEAO, LHNO, etc.)
- O Middelbaar beroepsonderwijs (MBO)
- O Hoger beroepsonderwijs (HBO)
- O Universitair onderwijs (WO Bachelor)
- O Hoger universitair onderwijs (WO Master)
- O Doctoraat, MBA, of vergelijkbaar
- O Anders, namelijk: _____

Vanaf dit punt is de input van de ander niet meer nodig.

De laatste vragen gaan over jouw eigen bereidheid om de programma's te kijken aan de hand van de gepresenteerde aanbevelingen.
We gebruiken hiervoor een interface gelijk aan die van NPO Start. Op het volgende scherm zie je hoe jouw drie persoonlijke aanbevelingen eruit zouden kunnen zien gebaseerd op de genre-opties die jij en je tv-partner zojuist hebben geselecteerd.

Geef aan in hoeverre **jij** het eens bent met de volgende stellingen, op basis van de persoonlijke aanbevelingen van jou en je tv-partner. Je kan de afbeelding nog eens bekijken door op de terug knop te klikken (pijl naar links).

	Helemaal oneens	Oneens	Neutraal	Eens	Helemaal eens
De aanbevolen programma's passen bij mijn interesses.	O	O	O	O	O
De aanbevolen programma's zijn nieuw voor mij.	O	O	O	O	O
Het aanbevelingssyteem heeft mij geholpen om nieuwe programma's te ontdekken.	O	O	O	O	O
De aanbevolen programma's zijn divers.	O	O	O	O	O

Daarnaast zijn we benieuwd naar je bereidheid om de aanbevolen programma's te kijken. Geef aan in hoeverre je het eens bent met de volgende stellingen.

Let op: er is ook een optie om aan te geven dat je het programma al een keer hebt gezien.

	Helemaal oneens	Oneens	Neutraal	Eens	Helemaal eens	Ik heb dit programma al een keer gezien
Ik zou de eerste aanbeveling kijken, wanneer de gelegenheid zich voordoet.	O	O	O	O	O	O
Ik zou de tweede aanbeveling kijken, wanneer de gelegenheid zich voordoet.	O	O	O	O	O	O
Ik zou de derde aanbeveling kijken, wanneer de gelegenheid zich voordoet.	O	O	O	O	O	O

Houdt voor de laatste vraag het volgende in gedachten:

Zodra dit aanbevelingssysteem zou worden geïmplementeerd in de NPO Start applicatie, zou ik...

	Helemaal oneens	Oneens	Neutraal	Eens	Helemaal eens
Dit aanbevelingssysteem opnieuw gebruiken.	O	O	O	O	O
Dit aanbevelingssysteem vaak gebruiken.	O	O	O	O	O
Mijn vrienden vertellen over dit aanbevelingssysteem.	O	O	O	O	O

Welke van de volgende programma's was onderdeel van de afbeelding die je zojuist bekeken hebt?

- O Ik vertrek
- O Draadstaal
- O Toren C
- O Nieuw zeer
- O De slimste mens
- O Jinek

Je hebt het einde van de vragenlijst bereikt! Hartelijk dank voor het invullen. Mocht je nog vragen en/of opmerkingen hebben kun je ze hieronder invullen.

Vergeet niet op het pijltje te drukken om je antwoorden op te slaan!