UNIL | Université de Lausanne
IDHEAP
Institut de hautes études
en administration publique

Stéphanie Würth

The effectiveness of the evaluation procedure at the Swiss National Science Foundation
The example of project funding in the division biology and medicine

Working paper de l'IDHEAP 4/2017

Stéphanie Würth

The effectiveness of the evaluation procedure at the Swiss National Science Foundation

The example of project funding in the division biology and medicine

MEMOIRE DE MASTER EN POLITIQUE ET MANAGEMENT PUBLICS (PMP)

Sous la direction du Prof. Martino Maggetti

et l'expertise du Dr. Martin Benninghoff

Mars 2017

UNIL

Faculté de droit, des sciences criminelles et d'administration publique

Institut de hautes études en administration publique (IDHEAP)

Working paper 4/2017

UNIL | Université de Lausanne

IDHEAP
Institut de hautes études
en administration publique

Table of contents

List of Figures

List of Tables

1 Introduction

1.1 Research and development in Switzerland

Switzerland is one of the most competitive countries with regards to research and innovation. In the European Innovation Scoreboard 2015 Switzerland remains the overall innovation leader outperforming the EU Member States (Hollanders, Es-Sadki, & Kanerva, 2015, p. 6). It is also among the countries spending the highest amount compared to the Gross Domestic Product (GDP) on Research and Development (R&D) activities. The private sector currently covers around two thirds of the R&D expenditures, which in 2012 amounted to CHF 16 billion. Public research funding in Switzerland mainly builds on researchers initiative, the principle of competition and international collaboration. Decisive for the allocation of funding is the quality of the submitted research proposals. Based on the Federal Act on the Promotion of Research and Innovation of 14 December 2012 (RIPA), the Confederation is responsible for the financing the promotion of research and innovation through the Commission for Technology and Innovation (CTI) and the Swiss National Science Foundation (SNSF), the latter of which will be the focus of the present study (SBFI, 2016).

1.2 Research funding at the SNSF

The SNSF was founded in 1952 as a private foundation in order to be independent. Administered by the Confederation, the SNSF funds basic science in all research disciplines. Over 3’400 projects, with more than 14’000 collaborators are funded each year. In 2015, an amount of CHF 878 million was allocated to the best researchers. It is thereby the most important research funding institution in Switzerland (SNSF, 2015d, p. 3, 2016c, p. 26).

The SNSF is divided into four main organizational units. The highest body, which takes the strategic decisions, is the Foundation Council and its Executive Committee. The National Research Council (NRC) is responsible for the evaluation of the applications submitted to the SNSF. Further, there are also Research Commissions at Higher Education institutions, which are locally based and act as a link to the SNSF. The Administrative Offices support the three organizational bodies mentioned above (SNSF, 2015d, p. 6).

Researchers can submit proposals in five funding categories: projects, careers, programmes, infrastructures and science communication. Project funding is the main funding scheme and accounts for approximately half of the SNSF’s total budget. This allows the researchers to determine the topic and goals of the projects independently. Hence creating a setting which allows for the pursuit of innovative ideas (SNSF, 2015d, pp. 5–10).

The submitted proposals undergo a competitive evaluation procedure. The NRC assesses the proposals and decides which projects are worth funding, taking into account external reviews provided by peers. The scientific quality of the proposed research project and the scientific qualifications of the researchers are examined. The evaluation procedure may slightly vary according to the funding category. In some schemes for example, career funding interviews with the applicants are carried out in the second stage of the evaluation procedure (SNSF, 2016e).

1.3 Relevance and research question

Many authors claim that peer review and therefore the evaluation procedure of many funding agencies is not effective, meaning that it does not lead to a selection of the best research projects. More specifically, these authors assume that peer review is conservative by nature, preventing novel approaches from being funded and therefore being biased against innovative and risky research (Berezin, 1998; Mitroff & Chubin, 1979; Rip, 2000; Wessely, 1998; Wood & Wessely, 2003). The present study aims to shed light on the question of the effectiveness of evaluation procedures building on peer review. The case of the SNSF is used because its evaluation procedure corresponds to the basic peer review process (cf. chapter 2.2.3 for a description), which makes it a valuable example. Evaluation studies confirm the high quality of the SNSF evaluation procedure while at the same time showing that some researchers believe that the SNSF may be biased against risky and innovative research projects (Coryn, Applegate, Schröter, Martens, & McCowen, 2012, p. 33; Langfeldt, Ramberg, & Gunnes, 2014, p. 75). This is in contrast to the SNSF's own statement that the approach sought in Project funding "(…) creates an environment where innovative ideas can be pursued" (SNSF, 2015d, p. 4).

Therefore, the research question, which will be addressed in this thesis is as follows: Do peer reviewers disadvantage innovative applications in the division biology and medicine? If so, is this due to the evaluation procedure implemented by the SNSF?

2 Literature review

In this chapter, I will examine the literature with relevance to the research question. It mainly focuses on principal-agent theory and peer review, followed by a short definition of the term innovation.

2.1 Principal-agent theory

In this section I will first outline the general theory. After that, the theory is applied to the context of science and research funding, as well as to the specific case of funding agencies. This first section is concluded by explanations on the position of the SNSF on the Swiss higher education, research and innovation sector.

2.1.1 General principal-agent theory

In a very broad sense, principal-agent theory describes situations in which one one party delegates work to another party, which then performs the work. The delegating party is called principal and the one executing the task is the agent (Eisenhardt, 1989, p. 58). Coleman (1990) gives a more detailed definition placing more emphasis on the exchange character:

"One actor who wants to accomplish a certain goal but lacks some of the skills or capacities necessary to do so finds another actor with those skills or capacities and obtains the latter's services in return for remuneration of some sort. The first actor may employ the second over a period of time, or he may merely contract for a particular service; there is a wide variety of patterns in which this type of relation is manifested. The general property which all these patterns have in common is that one actor (the second party in the above discussion) carries out actions (often directed toward a third party) which are intended to fulfil the interests of the first party" (Coleman, 1990, p. 146).

The first actor, the principal, yields his own actions and decisions for a defined and limited area, in turn gaining the right to control the actions and decisions of somebody else. The agent, as the second actor, gives up his right of control in a defined and limited area, in turn acquiring the right to act and decide for somebody else (Braun, 1993, p. 137).

Due to potentially conflicting goals between principal and agent and a possible information asymmetry, when the principal has less knowledge about how to achieve his objectives than the agent, two problems are likely to arise:

1) Moral hazard: the agent does not make an effort when performing the delegated task.

2) Adverse selection: the principal is not able to select the suitable agent, because the latter is misrepresenting his abilities and/or skills (Braun, 1993, p. 138; Braun & Guston, 2003, p. 304; Guston, 2000, p. 21).

The principal has several means to counteract these problems. In order to reduce the information asymmetry, which is related to both problems, the principal can invest in monitoring and reporting. The principal has the right to monitor because he transfers resources to the agent. However, monitoring means that the principal has to make an effort himself and therefore bear additional costs. Reporting in return, as long as it is self-reporting without further incentives, is unreliable (van der Meulen, 1998, p. 400). As far as information asymmetry concerns the lack of observation of the agent's behaviour, the principal can either try to observe the agent's behaviour by investing in information systems or make contracts based on the outcomes of the agent's behaviour (Eisenhardt, 1989, p. 61). This indicates that the costs to overcome an information asymmetry can be very high (Arrow, 1985, p. 48).
Concerning the conflicting goals the principal can try to strengthen the goal-sharing with the agent, for example by formulating, articulating and evaluating the pursuit of goals cooperatively and publicly (Guston, 1996, p. 231).

2.1.2 Principal-agent theory in the context of science and research funding

The fundamental problem of science policy is the asymmetry of information between the scientists who conduct research and the policy makers (politicians and administrators), who govern research. Scientists know a lot more about how to do research than policy makers. It is the scientists who know how to fulfil the objectives of the policy makers. This results in the difficulty for policy makers to monitor the researchers while at the same time researchers struggle to prove that they are actually making an effort (Guston, 1996, p. 230). But as researchers are producers and consumers of science simultaneously, they have an incentive to monitor each other. The process of peer review represents the institutionalization of this "professional consumer control" (van der Meulen, 1998, p. 400). Hence in order to overcome the asymmetry of information, the cooperation of scientists is necessary (Braun, 2003, p. 310). Researchers might not share the goals advocated by policy makers either. They possibly care more about their role and position in the scientific system, than about the policy makers' goals and interests. As a result they may prefer to do research on subjects different from the ones identified by policy makers (Guston, 2000, pp. 20–21; van der Meulen, 2003, p. 400).

Adverse selection

In the context of science policy, adverse selection concerns choosing the researcher who shares the policy maker's goals and is most suitable in order to realize their objectives. This concerns selecting the appropriate agent and it is a problem that occurs before a contract is made (Gulbrandsen, 2005, p. 200; Guston, 1996, p. 233, 2000, p. 21).

For the principal, it is difficult to pick the most suitable agent, because, as elaborated above, he lacks information about the agent (Guston, 2000, p. 21). In order to resolve this problem and select the appropriate agent, the principal often relies on the judgement of other agents. This implies a delegation and a review process, which in the end also benefits the researchers, as they rely on each other's results and the review process serves as a quality control. Their own future funding could be endangered were the research they selected irrelevant or of low quality (Borlaug, 2015, p. 3; Gulbrandsen, 2005, p. 205; Guston, 1996, p. 236). For more details about peer review and its possible disadvantages, see chapter 2.2.4.

Moral hazard

Moral hazard relates to the fact that "(…) the delegation by the principal provides not only an incentive to perform the required task, but also an incentive to cheat, shirk, or otherwise act unacceptably" (Guston, 2000, p. 21). This is a performance problem that only arises after the contract has been signed (Guston, 1996, p. 233). The scientists may be acting in a negligent or fraudulent way when conducting research or they may simply be pursuing goals other than those stated in the contract. Thus, the policy makers need to assure that the researchers do their best when conducting the research and that they do not act in unacceptable ways. The two predominant worries in this context are: making sure that research is conducted with integrity and that it is carried out productively (Guston, 2000, p. 23).

One way for the principal to deal with this problem is to invest in evaluation procedures coupled with output measuring indicators (Borlaug, 2015, p. 3).

This indicates that policy makers and scientists are in a delegation relationship. The policy maker, as the principal, requires the agent to perform specific tasks they are not capable of performing on their own, because, unlike the scientists, they do not have the capabilities and the knowledge (Guston, 2000, p. 15). Both actors depend on each other, and science policies need to balance the interests of the government as a principal and of science as a group of agents. Consequently, both will benefit from an intermediate structure that changes their preferences (van der Meulen, 1998, pp. 404, 412).

2.1.3 Funding agencies from the principal-agent perspective

Funding agencies[1] are created to develop and implement research policies. These are preferred to the common public administration, because they have indispensable direct contacts with science at their disposal, which will increase the likelihood of beneficial outcomes (Braun, 2006, p. 154; Braun & Guston, 2003, p. 303). Funding agencies can be seen as a link in an "iterated principal-agent relationship" (Guston, 1996, p. 231). The people principal to the policy makers are the first agents, who delegate authorities and install agencies embodying the next agents. The agency itself is the principal to the scientists (cf. Figure 1a). Following these considerations, a funding agency would be an agent and a principal at the same time (Guston, 1996, p. 231; van der Meulen, 2003, p. 324).

Fernández-Carro (2007, p. 323) pointed out that funding agencies can hardly be the scientists' principal, because it is not them rewarding scientists and they only command them very indirectly. In the same manner scientists are not simply agents of funding agencies, as they are autonomou and can influence decisions of the funding agencies. The same applies to the relationship between policy makers and funding agencies. This implies that the relationships are interdependent, that both parties offer the other something crucial and that autonomy is essential (Braun & Guston, 2003, p. 305).

Braun (1993) suggests that in a political system the principal-agent model should be considered as a triadic structure. As a consequence, he introduces a third party into the model. This results in the principal-agent relationship being between the government as the principal and the funding agency as the agent, with the scientist incorporating the third party (cf. Figure 1b) (Braun, 1993, pp. 136–141).

Van der Meulen (2003) considers funding agencies also to be intermediate bodies. But as opposed to Braun (1993), he does not consider them as the agents at the same time. In his configuration scientists embody the agents and the government remains the principal. Additionally, he brings conceptually different third parties into the game (cf. Figure 1c). Third parties in his view could, for example be industries or users. The intermediary can strengthen his strategic position if he focuses on the third party's interests as an approximation for the interests of the principal and the agent. The creation of an intermediary agency entails the transfer of authority and monitoring rights by the agents to the intermediary, which the agent in

[1] In some countries (e.g. Norway, United Kingdom) funding agencies are called research councils. In the case of the SNSF the National Research Council is the body responsible for the evaluation of the grant applications. Therefore, and for the sake of simplicity only the term funding agency is used to describe the whole organization responsible for the promotion of research.

the basic principal-agent relationship gives to the principal. In this case it is the funding agency that disposes of the monitoring rights. Its relative position to the policy makers and the scientists depends on the authority that is delegated, the funds it obtains by the government and to what extent scientists transfer monitoring rights to the agency (van der Meulen, 2003, pp. 324–326).

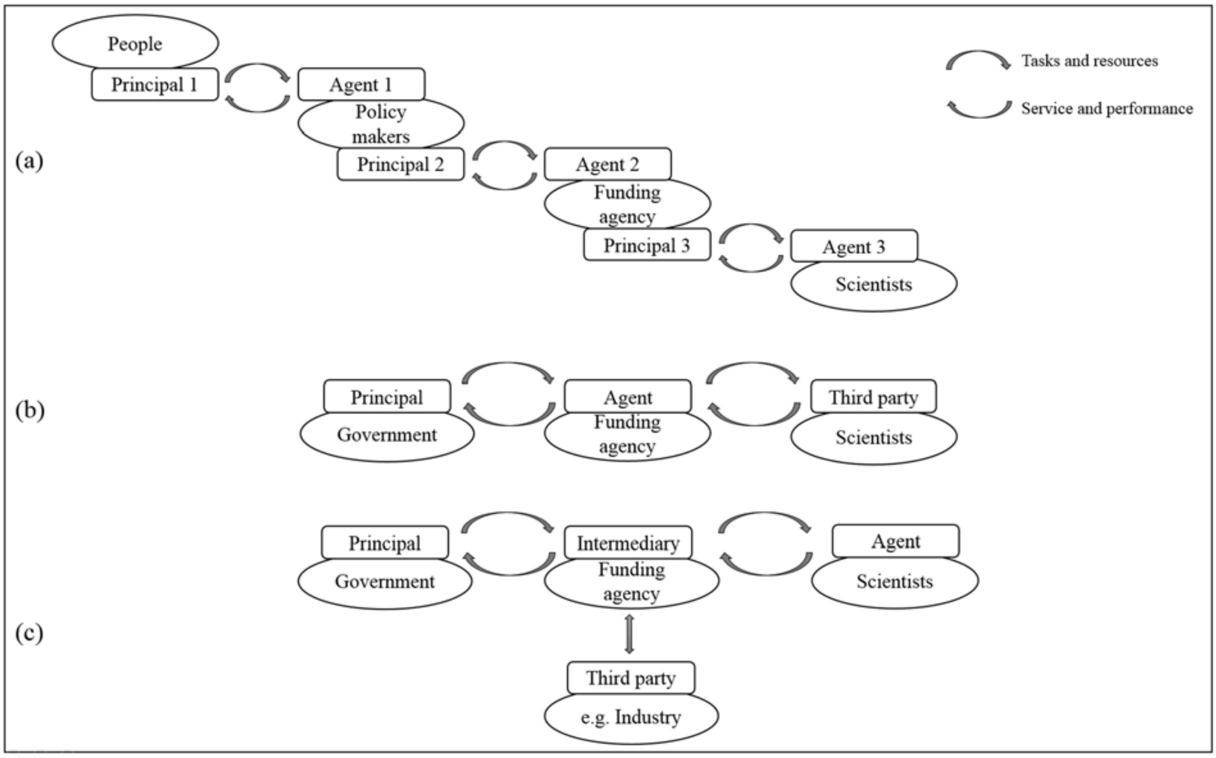

Figure 1. Principal agent configurations (authors own figure inspired by Braun (1993, p. 141) and van der Meulen (2003, p. 325))

A funding agency intermediates between policy makers and the scientists. It informs the scientists about the needs and interests, which are not its own but the policy makers'. In the same manner, the performances it informs the principal about are the scientists' (van der Meulen, 2003, p. 324). The scientists themselves are interested in collaborating with funding agencies, in order to influence the intermediaries' decisions and make it represent scientists' interests towards policy makers (Gilardi & Braun, 2002, p. 155). Thus, the intermediary agency's interests are determined by the interests of the other two actors and it helps to create and preserve the trust of policy makers regarding the scientists' work (van der Meulen, 2003, p. 324). A funding agency's major role is to fund research, but by choosing funding schemes, priority areas and evaluation procedures, it also plays a role in the development of research policy. Therefore the funding agency's success also depends on its ability to convince scientists to participate in funding schemes and to do their best to generate new knowledge. Towards the policy makers it is responsible for the implementation of policies. At the same time, however, it might pursue its own interests and try to increase the funds allocated to science, as their survival depends on money from the government (Braun, 2006, p. 161; Slipersæter, Lepori, & Dinges, 2007, p. 402).

Organisations, strategies and actions developed by funding agencies to satisfy policy makers and/or scientists, are dependent on the characteristics of the relationship with those two actors (Slipersæter et al., 2007, p. 403). A more thorough view of these relationships in the case of the SNSF is provided below.

2.1.4 SNSF’s position in the Swiss higher education, research and innovation (ERI) sector

Actors in the Swiss ERI-sector

The ERI-actors can be classified into three groups of actors. The first group consists of the suppliers of the necessary ERI-performances. They contribute to the education and generate new knowledge and innovations. Among these are the institutions of vocational and professional training, the different types of universities, research facilities of national importance and private R&D companies.

Regarding the intermediary functions, the SNSF is not the only actor in the Swiss ERI-sector. The CTI, the academies, the Swiss Science and Innovation Council (SSIC), swissuniversities the Rectors' Conference of Swiss Higher Education Institutions as well as the Board of the two Federal Institutes of Technology (ETH Board) and the Swiss Accreditation Council also perform intermediary tasks.

Among the policy makers who take decisions with relevance to the ERI-sector on the federal and cantonal level, are the Science, Education and Culture Committees (SECC) of the two chambers of the Federal Assembly, the Federal Department of Economic Affairs, Education and Research (EAER) or rather the State Secretariat for Education Research and Innovation (SERI), the Swiss University Conference (SHK), the Conference of Cantonal Ministers of Education (EDK) as well as the cantonal ministers of education and of finance. It shall be noted that in the case of the SNSF the Cantons have no regulatory competence. This competence remains solely in the hands of the Confederation (SSIC, 2015, pp. 17, 22).

The functions and instruments of the actors promoting research and innovation on the federal level are defined in the RIPA. The SNSF intervenes at different phases of the cycle consisting of policy formulation, consultation, coordination and evaluation in preparation of decision making. It plays a part in the definition of policy on the federal level and has an advisory function, together with the ETH Board, the SSIC and the academies. The SNSF also decides about promotion of research in case of project funding. Finally, it also helps to evaluate the implementation of science policy (SSIC, 2015, pp. 20–21).

SNSF's responsiveness towards policy makers

Slipersæter, Lepori and Dinges (2007, p. 405) found that responsiveness of funding agencies towards policy makers depends on various factors: dependency on the number of funding sources, the type of agency, its mission, the characteristics of the circle of beneficiaries and the constitution of the funding agency's board. In the following paragraphs, these factors will be analysed with regard to the SNSF.

The vast majority of the SNSF' funds come from the Confederation (SNSF, 2016c, p. 30). The payment structure is authorised by the Federal Assembly for a multi-year period (FIFG, 2014, art. 36 let. a). It would be at this point that the government has the most valuable opportunity to affect the SNSF's policy and set particular objectives (Slipersæter et al., 2007, p. 408). However, according to the analysis of the SSIC (2016) the Confederation, as well as the Federal Parliament in Switzerland act mainly as a hub of resource allocation and decentralize the risk and responsibility involved in strategic decisions. The conviction that the government should respect funding agencies' autonomy seems to prevail (SSIC, 2015, p. 29).

The SNSF belongs to the group of "all-purpose agencies", which mainly fund basic research and are not limited to a certain discipline or research area (Braun, 2006, p. 159). The SNSF is bound by law to place particular emphasis on the promotion of basic research and to fund research in all academic disciplines (RIPA, 2015, art. 10 par. 1 & art. 9 par. 4). According to Braun (2006), this type of funding agency disposes of very substantial discretion (e.g. the right to decide and develop policies) because uncertainty about preferences, objectives and outcomes is high for policy makers in basic research. Control procedures in order to prevent and/or mitigate the shirking of funding agencies on the other hand are extremely limited. As far as shirking concerns respecting policy preferences this limitation is not a problem, because in basic research the policy makers seldom have precise preferences (Braun, 2006, pp. 155, 159, 166–167).

The main aim and SNSF's priority is the responsive funding mode based on competition between the applicants. Most of the available money is attributed to project and career funding, without any conditions regarding subjects, strategy or structure (SNSF, 2015c, p. 4). One of SNSF's strategic goals is to align research funding with the researchers' needs, indicating that the SNSF tries to create the best possible environment for scientists, which is needed for the promotion of basic research (Braun, 2006, p. 163; SNSF, 2016b). Still, an all-purpose agency like the SNSF has to respond to policy makers' requests, especially in the case of applied research, even more so because it depends on one single funding source. The SNSF first fulfilled such a demand by implementing National Research Programmes (NRP) aimed at solving "Switzerland's most pressing problems" in the 1970s (Braun, 2006, pp. 163, 165; Slipersæter et al., 2007, p. 413; SNSF, 2011b, p. 2). NRPs together with National Centres of Competence in Research (NCCR) are one of the few ways that the Confederation can influence the higher education sector by participating in the decision of research themes and inducing long-term structural impact (Slipersæter et al., 2007, p. 413; SSIC, 2015, p. 24). However, in 2015 only about 9% of the grants awarded went to the two thematic instruments NRP and NCCR (SNSF, 2016d). This illustrates the fact that the Confederation in general continues to follow a strategy of renouncing the pursuit of strategic goals (SSIC, 2015, p. 29). Moreover, in the last twelve years expenditures in the ERI-sector increased more than in other policy areas (SSIC, 2015, p. 28). For the SNSF the yearly growth amounted to 3.5% between 2013 and 2016 (SNSF, 2013, p. 7).

The SNSFs evaluation body, the NRC, is composed of a maximum of 100 researchers, who are elected by the Executive Committee of the Foundation Council for four years. The Foundation Council is composed of members from scientific organisations and members from business and politics appointed by the Federal Council. Fifteen members thereof form the Executive Committee (SNSF, 2002, art. 8, 9, 12 par. 2 let. a, 14, 18, 19). According to Slipersæter et al. (2007, p. 410) it is the NRC together with the Administrative Offices that adopts most strategic decisions and documents. This points to the tendency that the scientific community indeed is able to influence the SNSF's decisions and that scientists' interests are well represented with regard to policy makers. At the same time scientists transfer monitoring rights to the SNSF, although these mainly remain in the hands of the community as it is the NRC that "(…)monitors and supervises the research work supported by the SNSF as well as the implementation of the corresponding research results" (SNSF, 2002, p. art. 21 par. 2 let. f).

According to Slipersæter et al. (2007, p. 413) the SNSF was able to defend its existence and strategies as well as the community's interests due to a rather weak government and an independent council that is responsive to and coordinated with a powerful scientific community. It should however, be noted that ERI-policy in Switzerland traditionnally and consistently follows a subsidiarity and self-regulation principle, which also explains the SNSF's success in maintaining its position (SSIC, 2015, p. 29).

2.2 Peer review

This section is introduced by general remarks on peer review. This is followed by the elaboration of reasons for the implementation of peer review in funding agencies. Then the general process of peer review, as well as the evaluation criteria are described. The ensuing subsection addresses the shortcomings of peer review and the corresponding empirical evidence. This section concludes with a description of the evaluation procedure of the SNSF's project funding and the evaluation criteria that are used.

2.2.1 Introduction to peer review

Peer review is the examination and evaluation by "equals", which in a professional context are usually expert colleagues (Gutknecht-Gmeiner, 2008, pp. 51–52). In science, peer review is carried out to assess the quality of colleagues' scientific work. It has two functions: firstly, it evaluates the quality of research proposals prospectively, therefore, determining which research obtains funding and secondly, through ex post evaluation of the quality of the research conducted, it determines which results get published (Bornmann, 2011, p. 199; Hartmann & Neidhardt, 1990, p. 419). The difference between grant and journal peer review is that even bad papers will get published somewhere, whereas grants that are not funded might be science that is not performed at all (Spier, 2002, p. 103; Wessely, 1998, p. 301). Scientists who perform this task should in any case be at the front of their research areas and be recognized as having the expertise required to judge the quality of research proposals (Wood & Wessely, 2003, pos. 291). During the last fifty years the importance of and reliance on peer review has increased tremendously as the number of people working in science grew substantially. This lead to enhanced competition for limited funds and space in scientific journals (Bornmann, 2011, p. 200; Spier, 2002, p. 358).

2.2.2 Reasons for the implementation of peer review in funding agencies

As has already been explained in chapter 2.1.2, funding agencies implement peer review to reduce the information asymmetry and therefore the adverse selection problem (Guston, 1996, p. 231). In this section the aim is to shed more light on the interests the two actors, scientists and policy makers, have in this process.

From the principal agent perspective, the scientists as the agents are in favour of peer review because this allows them to remain in control of monitoring and to have discretion over funding allocation. The principal uses the outcome of peer review to make decisions about allocation and to define new strategies. Additionally, it is also a monitoring instrument to make sure that funds are allocated to the appropriate agents (van der Meulen, 1998, p. 405).

From a more sociological point of view, peer review serves the interests of the scientific community and the policy makers, because it legitimates funding decisions to both of them. König (2015, p. 3) calls this the "dual legitimacy", which only peer review is able to generate as a decision-making principle in research funding. Dual legitimacy means that peer review is able to meet the expectations of both the scientists and policy makers at the same time. Scientists are primarily concerned with the fairness of the decision-making, whereas policy makers are primarily interested in its efficiency. For the procedure to be fair, the process needs to be defined clearly. It must also ensure that all proposals submitted are handled equally and it has to be clear that the reviewers themselves are impartial. The procedure then results in a fair, scientific judgement of the grant application's quality. Efficiency requires that the procedure is outsourced to another organizational entity and that evaluation criteria are revealed (König, 2015, pp. 9–12). These requirements coincide to a large degree with the process characteristics Chubin (1994) identified as desirable. Two characteristics not explicitly referred to by König (2015) are worth noting: the process should also be effective, meaning that the research targeted by a funding scheme should be supported; further the process needs to be responsive, implying that policy makers can guide research efforts and promote emerging areas. Another important point Chubin (1994) noted is that trade-offs between objectives are inevitable and that compromises among them will have to be made (Chubin, 1994, pp. 23–26).

2.2.3 Process of grant peer review

Three steps can be identified in the basic process of peer review.

1. *Screening of proposals*

First, employees of the funding agency review submitted proposals. They check whether the applicants are eligible and whether the proposals correspond with the formal requirements (Sharif, Farrands, & Wooding, 2009, p. 5).

2. *Review of proposals*

In general, two groups - external and internal reviewers - are involved in this step. External reviewers generally evaluate only one proposal per call and do so independently from other reviewers. Internal reviewers belong to a review committee that will discuss and assess a group of proposals. This discussion builds on the assessments obtained from the external reviewers and leads to a ranking of the proposals, which is based on comprehensible criteria. There are two major decision making modes a review committee can employ: unanimous agreement or consensus by majority (ESF, 2011, pp. 23, 32).

The selection of external reviewers is crucial. It is advisable that there is sufficient distance between the external and internal reviewers. Usually the employees choose the external reviewers, although members of the review committee as well as applicants can suggest reviewers to be included or excluded too. The selected external reviewers should meet certain criteria like scientific excellence in the respective field, expertise in reviewing, independence from the funding agency and no conflict of interest (ESF, 2011, pp. 25–26).

3. *Final decision*

Usually a board distinct from the review committee takes a decision based on the outcomes of step two and applicants are informed accordingly. It is recommended that there are at least three reviews before the decision is taken (ESF, 2011, pp. 25, 33; Sharif et al., 2009, p. 6).

Evaluation criteria

The European Survey on Peer Review Practices identified four groups of evaluation criteria: relevance and expected impacts, scientific quality, qualification of the applicant(s) and availability and suitability of the research environment (ESF, 2011, pp. 28–29).

Using quantitative content analysis Neidhardt (1988) identified 11 categories of criteria actually used by internal and external reviewers in the evaluation of proposals at the Deutsche

Forschungsgemeinschaft (DFG)[2]. These categories read as follows: qualification / reputation, preparatory work, scientific significance, practical significance, theory, method, feasibility, planning, budget, unspecific comments and miscellaneous. Arguments concerning the categories' theoretical quality, scientific significance, appropriateness of the budget and qualification and reputation of the principal investigator were mentioned the most often - in around 42-45% of all examined reviews. Less often, in about 30% of the reviews, were comments in the categories planning, feasibility of the project and its practical significance as well as unspecific comments. The reviewers commented positively on most of the criteria. All the categories of criteria were found to have a significant impact on the funding recommendations. This indicates that not a single criteria, but rather a constellation of arguments determined the recommendation with the categories budget, theory, method and qualification / reputation having the biggest influence (Hartmann & Neidhardt, 1990, pp. 420–424; Neidhardt, 1988, p. 101).

Reinhart (2012) performed a quantitative content analysis, similar to the one carried out by Neidhardt (1988) using data on all proposals treated by the SNSF in the division of biology and medicine in the year 1998[3]. He found that the comments on priority and summarizing appeared most often, followed by comments on methods and originality, which were also among the most frequently mentioned criteria in the study from Neidhardt (1988). Overall most of the comments (63%) were positive; few were negative (24%) and even fewer neutral (13%). As in Neidhardt's study, criteria related to the researcher (topicality, environment, reputation, practical relevance, qualification, and co- applicant) are more often positively assessed, while criteria referring to the project (presentation, research plan, methods, and costs) are more frequently associated with negative statements. More than half of the examined criteria appeared in more than 50% of the reviews written for the SNSF, whereas in the DFG study no criteria was mentioned in more than 45% of the reviews. So, external reviewers for the SNSF used a greater variety of quality criteria in their reviews than reviewers for the DFG. This might be explained by the fact that the DFG study mainly analysed reviews from internal reviewers, who probably need to state less criteria explicitly in order to justify their evaluation (Reinhart, 2012, pp. 161–170).

[2] The DFG, at the time, did not provide the reviewers with evaluation guidelines, defining criteria to be used in the assessment. It was up to them to specify and weigh criteria (Neidhardt, 1988, p. 86).

[3] At this time potential external reviewers were not systematically provided with detailed instructions, as is the case now. Two versions of accompanying letters were in use, with only the longer one explaining the evaluation criteria to be employed. In retrospect, however, it was impossible to deduce, which letter version was sent to whom.

Analysing the review guidelines of fourteen calls for proposals in medical research a study found very similar criteria. Ensuing interviews with internal and external reviewers showed that they perceived originality and methodology as most important, followed by scientific relevance and feasibility (Abdoul et al., 2012, pp. 3, 11–12).

A study among funding panels in the United States and Canada found that when reviewers discuss originality they use various kinds of arguments. They frequently concentrate on the aspects of the proposal itself, but also consider the character of the applicant or their own emotions regarding the proposal (Lamont, Fournier, Guetzkow, Mallard, & Bernier, 2007, p. 177).

This demonstrates that even though the quality of a project is important when decisions about funding are made, it is not a sole criterion. In reality many different criteria play a role. Moreover, it seems that guidelines only moderately influence the criteria emphasized by internal reviewers, whereas external reviewers seem to try harder to use the guidelines for their reviews (Langfeldt, 2001, p. 835).

2.2.4 Shortcomings of peer review

Even though peer review is widely used, it is still subject to a lot of criticism. Drawing on the work of Bornmann (2011), Wood and Wessely (2003) and Sharif, Farrands and Wooding (2009) criticism can be divided into five main categories.

1) Biased recommendations: reviewers’ recommendations are not always based on scientific quality only. They are also influenced by the characteristics of the applicants or reviewers themselves. This would imply that fairness of peer review is not a given although, as seen above, it is essential to legitimate funding decisions towards the scientists.
2) Inefficiency: peer review creates inefficiencies because it consumes a lot of time and is costly. Such inefficiency would impair the legitimation policy makers assign to peer review.
3) Ineffectiveness: peer review does not fund the best science, meaning that, for example, the promotion of innovative research is restricted[4].

[4] It is worth noting that the definitions of efficiency and effectiveness vary between authors and stakeholder groups. Some include reliability in the discussion around the effectiveness, whereas others subsume effectiveness and efficiency under the term efficiency. Here efficiency refers to time and money and effectiveness to the selection of the best science.

4) Poor reliability: reviewers seldom agree on whether a grant should be recommended for funding.
5) Lack of predictive validity: reviewers' judgements are only weakly related to the work's future value for the scientific community (Bornmann, 2011, pp. 203–204; Sharif et al., 2009, p. 8; Wood & Wessely, 2003).

Empirical evidence from research on peer review

This section will focus mainly on empirical evidence for the effectiveness of peer review, as the research question is concerned with this potential problem. Evidence regarding the other four possible biases will be summarised briefly.

Biased recommendations

Gender is one of the most studied possible biases. The numerous studies came to different conclusions. A comprehensive meta-analysis showed that there is no evidence for a gender effect disadvantaging women (Marsh, Bornmann, Mutz, Daniel, & O 'mara, 2009, p. 1311).
Neither is the alleged bias against some institutions supported by research. However, one study showed that reviewers tend to favour grant applications that back their own school of thought (Sharif et al., 2009, p. 17).
It is not clear either, whether peer review disadvantages young scientists. The results of the existing studies are contradicting. As experience is also influenced by age, the interpretation of findings is further complicated (Sharif et al., 2009, p. 17; Wood & Wessely, 2003, pos. 378–385).
Thus, it can be concluded that it is not clear whether the criticism of peer review being biased is valid or not. Even if some biases were existent, it is not clear how they would impact on research quality (Wood & Wessely, 2003, pos. 594–601).

Inefficiency

The time spent on the writing and reviewing of grants is substantial. Overall the number of applications, as well as of re-submitted applications is increasing. The acceptance rates for reviews are decreasing, indicating that for funding agencies it is increasingly difficult to find scientists who are willing to give up time to review. This raises the funding agencies' administrative burden, as they have to contact more potential reviewers. In combination with decreasing success rates, a growing inefficiency in the peer review process seems possible (Sharif et al., 2009, pp. 9–13; Wood & Wessely, 2003, pos. 490–499).

Ineffectiveness

There seems to be general agreement that peer review serves well when it comes to preventing the wastage of resources on poor science. It also helps to reduce the risk of spending money on promising but not feasible research proposals (Rip, 2000, p. 468; Wood & Wessely, 2003, pos. 291). But the really important question regarding the effectiveness of peer review is, whether innovative research proposals receive funding and scientists are able to make important discoveries (Horrobin, 1996, p. 1293; Wood & Wessely, 2003, pos. 610–618). There are numerous authors who assume that peer review is inherently conservative, discouraging novel approaches and therefore biased against risk-taking and innovation (Berezin, 1998, p. 5; Mitroff & Chubin, 1979, p. 209; Rip, 2000, pp. 468, 472; Wessely, 1998, p. 303; Wood & Wessely, 2003). The existing empirical evidence however, is considered insufficient to support these claims (Sharif et al., 2009; Wood & Wessely, 2003).

The researchers themselves also do not seem to be that negative about the funding agencies' ability to promote innovative research. Two surveys performed among applicants in Germany at the DFG and at the US National Science foundation (NSF) found that only few researchers thought that decisions are biased against innovative ideas (Böhmer, Neufeld, Hinze, Klode, & Hornbostel, 2011, pp. 77–78; McCullough, 1989, pp. 81–82). Researchers in Austria think that originality and a level of innovation are already given a rather high weight in review, but wish that these factors would be given even more weight. They would prefer that the fact of belonging to the actual mainstream research be given much less weight than it is given now (Neufeld, Hinze, & Hornbostel, 2014, pp. 31–32).

The following few studies investigated empirically whether peer review is effective. Horrobin (1990) came to the conclusion that the trade off between innovation and quality control, among other factors, was responsible for the fact that fewer improvements had been made in patient care since 1960. Yet, his analysis only relied on a limited number of examples, showing how peer review resulted in judgement against innovation (Horrobin, 1990, pp. 1439–1441).

Based on interviews with members of the European Research Council (ERC) panels Luukkonen (2012) concluded that "(…) the peer review process in some ways constrains the promotion of truly innovative research. (…) However, this does not necessarily mean that peer review prevents new openings, especially if such an aim is a central evaluation criterion" (Luukkonen, 2012, p. 58).

Using an experimental research design another study showed that more novel proposals get lower grades. Responsible for this result are proposals with especially high levels of novelty.

For proposals with low levels of novelty the grades increased with degree of novelty (Boudreau, Guinan, Lakhani, & Riedl, 2016, p. 2).
Interviews with researchers from Germany and Australia revealed that they adapt their strategies to the funding conditions, with one of the strategies being the avoidance of risky research. If scientists are not sure whether a new idea will work, they fear that they will not obtain funding and refrain from including it in their grant application (Laudel, 2006, p. 497).

Reliability

Existing studies produced different results regarding the reliability of peer review. For dichotomous decisions, funded or not funded, reliability is usually higher than for the ratings of applications. Furthermore, reliability is greater for grants of poor quality than good quality grants. Overall, it can be noted that the peer review is not perfectly reliable and that some randomness prevails, but this is mainly due to a the absence of consensus in fields on the frontiers of knowledge and different, normatively correct interpretations (Lee, Sugimoto, Zhang, & Cronin, 2013, p. 6; Sharif et al., 2009, pp. 15–16; Wood & Wessely, 2003, pos. 439–447, 601).

Predictive validity

Regarding the predictive validity of peer review, results are also mixed. One study came to the conclusion that ratings attributed by the review process seemed to be linked to the subsequent performance. Although this might be explained by the fact that better rated applications are usually awarded higher budgets and a longer duration. A more recent study using data on grants from another funding agency found that the rating had no influence on the number of publications and citations (Mervis, 2014, p. 596; Wood & Wessely, 2003, pos. 469). Several studies demonstrated that research accepted for funding by peer review usually has more impact than unfunded research. But there are no studies determining the effect that peer review has on the quality of the funded research (Demicheli & Pietrantonj, 2007, pp. 3–4; Wood & Wessely, 2003, pos. 469–474).
Overall it seems that the empirical evidence is indeed limited and it seems plausible that some of the assumed shortcomings of peer review emerged on the grounds of individual knowledge and non-objective analysis (Demicheli & Pietrantonj, 2007, p. 6).

2.2.5 Evaluation procedure at the SNSF in the case of project funding

The evaluation procedure for project funding, the SNSF's principal funding scheme, is clearly defined and explained on its website. It generally lasts six months and is divided into three successive parts, coinciding with the three stages of the basic peer review process:

1. *Submission of applications and administrative measures*

The applications are to be submitted online via the mySNF platform by the submission deadline. The Administrative Offices of the SNSF check whether the conditions for submitting an application are fulfilled and inform the applicants. In case of shortcomings the Administrative Offices either set a deadline for these to be corrected or do not consider the application. Based on the discipline or sub-discipline and the keywords, applications are assigned to a specific evaluation body of the NRC. Further, a referee plus a co-referee, from among the members of the NRC, are assigned to each project based on expertise and workload. In a first step they examine whether a reason for direct refusal is present. If so, the application is not sent out for external review (SNSF, 2011a, pp. 2–4).

2. *Evaluation*

The evaluation is a two-stage procedure. Firstly, external reviewers evaluate the applications online via the mySNF platform in accordance with SNSF standards[5]. They assess the proposals according to the evaluation criteria that are further specified in guidelines put at their disposal and give a rating from "outstanding" to "poor" for each criterion. In the end they also give a justified assessment of the overall quality of the project. The external assessments are made available to the applicants once the final decision has been made, whereof external reviewers are informed beforehand (SNSF, 2016e).

The referees usually decide which persons are to be asked for an external review. The Administrative Offices help the referees to find suitable reviewers by providing a list of potential external reviewers. The applicants have the right to provide a list of persons who shall not be asked for a review. The SNSF respects this list as long as the proposed exclusion is well justified and enough other reviewers are available (SNSF, 2007, art. 23, 2015c, art. 25). Until 2016 they could also suggest researchers to be asked to review their application. The SNSF carries out extensive checks to ensure that the reviewers do not have conflicts of interest and that reviews are impartial. In principle, two external reviews per application are

[5] If there are a sufficient number of easily comparable applications, the SNSF can also set up a reader system or a panel for the purpose of peer review.

required (SNSF, 2015c, art. 25 par. 2). Usually this requirement is met, as the SNSF on average receives three external reviews, whereby international experts do most of them. But the review return rate decreased over time. Thus the number of invitations sent out to potential external reviewers had to be increased (SNSF, 2015a, p. 2).

In the second stage, the responsible referees of the NRC assess the applications on the basis of the external reviews. They assess whether the received external criteria-based reviews are useful and critically comment and complete them using the same criteria as the external reviewers. Then they rate the application in relation to other applications in their responsibility, on a scale from one to six. This rating needs to be justified, based on a short description of the strengths and weaknesses of the application in question. For applications likely to be funded, they make a financing proposal. Finally, they issue a recommendation to the relevant evaluation body to adopt their rating (SNSF, 2016e).

As the identity of external reviewers and referees remains concealed throughout the procedure, whereas the applicant's identity is revealed to reviewers and referees, the evaluation procedure at the SNSF belongs to the group of single-blinded assessments (ESF, 2011, p. 14; SNSF, 2016e).

3. Decision

The relevant evaluation body debates the applications comparatively after a brief presentation of arguments by the referee and co-referee. Afterwards each application is assigned to one of the six categories (distribution-based assessment) and gets either approved or rejected by majority vote. Usually the final grades given by the NRC are lower than the grades by the external reviewers, who in general write more positive reviews. This is mainly due to the fact, that external reviewers normally only assess one application, whereas the SNSF evaluation body has to consider the quality of all the submitted applications (SNSF, 2015a, pp. 7–8).

In a final meeting the available budget is distributed among the highest-rated applications. As a result, a list with provisional approvals, including financing proposals and rejections is adopted. This list is passed on to the Presidency of the NRC that examines the correctness of the procedures, adherence to the budget and compliance with other conditions. If these are approved, the provisional decision is endorsed en bloc and made final.

The applicants are informed by the SNSF of the final decision by the means of a ruling about six months after the submission deadline. In case of rejection, a justification for the decision is included (SNSF, 2016e).

Evaluation criteria and principles

The following principles serve as a basis for the rules and practices of the evaluation and decision-making of the SNSF: excellence through competition, fairness and equal opportunities, transparency, integrity and confidentiality (SNSF, 2016e).

The evaluation criteria, which are applied by external reviewers and referees, are broadly defined in the Regulations of the SNSF on research grants. Article 24 states:

„The scientific evaluation is based on the following main criteria:

a. scientific quality of the proposed research project: scientific relevance, topicality and originality, suitability of methods, feasibility;
b. scientific qualifications of the researchers: scientific track record and ability to carry out the research project“ (SNSF, 2015e, p. 9)

The guidelines for the assessment give a more thorough definition of what is meant by these criteria (SNSF, 2014):

- Scientific relevance: The scientific relevance includes the relevance of the topic and the research problems or hypotheses, the projects' potential to increase the knowledge and coherence and the project's ability to develop the scientific approaches and methods.
- Topicality: Proposals introducing subjects of current interest are considered topical. The number of recent publications and citations as well as references to relevant recent events may indicate topicality.
- Originality: Originality of the research problem and the theoretical / methodical approach proposed in the project. A proposal introducing questions that so far have been neglected or an approach that combines known aspects in a novel way, may be considered as original.
- Broader impact: This criterion is only to be assessed in cases where a project is classified as use-inspired research. The assessment looks to what extent the proposed project has a broader impact, whether practitioners perceive a need for research and whether the results can be put into practice.
- Suitability of methods: This criterion relates to the extent to which the methods are suited to answer the research questions proposed in the application. It concerns the choice of methods, their combination and the coherence of the research plan.
- Feasibility: To what extent is the proposed research project feasible. The probability that the proposed milestones can be reached in the given time, with the available financial and personnel resources is to be considered for this purpose.

- Applicants' scientific track record: Publications, preferably peer reviewed, but also patents as well as science communication and networking activities shall be acknowledged in the assessment of this criterion. The scientific work is to be considered relevant the more it lead to progress within the discipline and beyond. Prizes, awards and citations are considered indicators for relevance. Therefore CVs and publication lists, from the last five years form the basis for the assessment of this criterion.
- Applicants' ability to carry out the research project: The applicant's or the team's expertise to carry out the proposed project successfully. Here the current state of the applicants' research specified in their research plan is to be considered too.

These criteria are used by the external reviewers as well as by the referees. In addition to the overall assessment, the reviewers attribute grades to three blocks of criteria: applicant's scientific track record and expertise; scientific relevance, originality, topicality (and broader impact) and suitability of methods and feasibility. The referees however, apart from the comparative ranking only grade the overall scientific quality of the proposed research project and the scientific qualifications of the researchers. Moreover, they can refer to the external reviews in cases where it would be useful and the judgements match (SNSF, 2011a, p. 7).
The study of Reinhart (2012, see chapter 2.2.3 above) showed that the criteria methods and originality were invoked in most of the reviews (in 66% of the reviews each)[6]. These are followed by feasibility (55%), scientific relevance (53%, called theoretical relevance in his study) and the applicants' ability (52%, called qualification in his study). Topicality (45%) and applicants' scientific track record (42%, called reputation in his study) appear less often.
Of these seven criteria, methods and originality were most often assessed negatively (35% of comments and 27% respectively), still 51% and 67% of the comments respectively were positive. Comments on topicality were almost always positive (95%). The scientific qualifications of the researchers were also often assessed positively (86% and 81%). The remaining categories, feasibility (62%) and scientific relevance (77%), also obtained a considerable amount of positive comments (Reinhart, 2012, p. 168).

[6] In total, Reinhart (2012) included more criteria in his study than the seven criteria that should be used in the scientific evaluation of the SNSF.

2.3 The concept of innovation in research

In literature, many different expressions are used as synonyms for and to describe innovation in research or innovative research projects. Among these are the following terms: novelty, originality, creativity, risky, high-risk / high-gain, frontier research, breakthrough and ground-breaking research. Their meaning and link to innovation will be elaborated below.

Innovation as such can be defined as follows:

"Innovation is the process of making changes, large and small, radical and incremental, to products, processes, and services that results in the introduction of something new for the organization that adds value to customers and contributes to the knowledge store of the organization" (O'Sullivan & Dooley, 2009, p. 5).

There are various types of novelty, such as in theory, research questions, application contexts, methods or statistical analyses. However, it is impossible to say that something is truly novel, as it cannot be excluded with certainty that somewhere in the world somebody has not already attempted to do the same thing. According to this definition innovation goes beyond novelty. For research to be innovative it has to be more than just doing something that has not been done before. Additionally, it requires that some value be added, so that it is useful and can be put into practice. This is what usually is referred to in research as impact (C. J. Lee et al., 2013, p. 10; Luukkonen, 2012, p. 54; Sternberg, Pretz, & Kaufman, 2003, p. 158).

In order to add value to something novel and to be useful, originality seems to be required (The Guidelines project, 2016). Originality in science means to work on something nobody else has worked on before and to advance scientific knowledge, hence, being useful to the scientific community. However, a precise definition, let alone a measure for originality in science does not exist (Dirk, 1999, p. 765; Lamont et al., 2007, pp. 169–171). Both, originality and novelty are prerequisites for creativity, which consequently is also in a close relationship with innovation. Depending on the definition of creativity, another component not contained in the concepts of originality and novelty has to be entered. This is surprise or non-obviousness (Simonton, 2012, pp. 97–99). However, creativity is also not to be equated with innovation as "innovation encourages the further processing of the output of the creative process (the idea) so as to allow the exploitation of its potential value through development" (O'Sullivan & Dooley, 2009, p. 7; 9), meaning that innovation takes creativity further.

Also inherent to novelty and therefore to innovation is the risk of failure. The more radical an innovation, meaning that major changes are made in something established, the higher the level of risk and the potential gain (Boudreau et al., 2016, p. 6; O'Sullivan & Dooley, 2009, p. 24;

Schumpeter, 1939, p. 102). If only safe projects were funded, this in itself would constitute a risk because innovation would be suppressed (Leung & Isaacs, 2008, p. 511).
From an economic point of view, risk means that probabilities can be assigned to different possible outcomes. The decision theory defines risk in a similar way. It describes decisions taken knowing the probability of the different possible outcomes. From the perspective of project management, risk is seen as something uncertain, which in case of occurrence has either a positive or a negative impact on the realisation of project objectives (Juite Wang, Lin, & Huang, 2010, p. 602). Generally a risk event is characterised by two elements: the likelihood of it occurring and the effect it may have (Baccarini & Melville, 2011, p. 222; Y. Lee, Chung, & Kim, 2007, p. 508). In research there is no way of knowing the probability of the possible outcomes, which might be undesirable. Risk is always associated with uncertainties arising from different sources (Luppino, Hosseini, & Rameezdeen, 2014, pp. 68–69; Merkhofer, 1987, p. 2).

The terms frontier, breakthrough and ground-breaking research, are usually used as synonyms; with frontier research being a concept that is politically loaded, since it was introduced by the European Commission in 2005. All these terms refer to research that reaches beyond the existing borders of knowledge, which is usually promoted through specific funding schemes established for this purpose. Innovative research funded through the "normal" instruments in return is thought to lead to more incremental results (ESF, 2011, p. 45; Luukkonen, 2012, pp. 54, 59; Scherngell et al., 2013, p. 249).

3 Hypotheses

The hypotheses are deduced from the literature presented above.

Hypothesis 1a: The more an application is perceived as innovative by an external reviewer, the lower the grade given by the reviewer.
Hypothesis 1b: The more an application is perceived as innovative by a referee, the lower the grade given by the referee.

Many authors claim that peer review discriminates against innovative research proposals, which would imply that peers in case of innovative proposals are not able or not willing to reduce the information asymmetry between principal and agent effectively. So far only one study empirically showed that novel proposals get lower ratings from reviewers (Boudreau et al., 2016). The study from Neidhardt (1988) examined what criteria have an influence on the rating, but did not pay specific attention to the degree of innovation of research proposals (Neidhardt, 1988).

In literature various explanations for the assumed conservatism are given and analysed.

- According to the theory of bounded rationality, peers are not capable of fulfilling their task to reduce the information asymmetry between the funding agency and the researcher. Bounded rationality describes how decisions are reached and under what conditions these decision-making strategies will fail or succeed (Gigerenzer & Selten, 2002, p. 4). It suggests that experts who have to go beyond the existing borders of knowledge when evaluating new ideas are likely to make systematic errors and inaccurately rate innovative research proposals (Boudreau et al., 2016, p. 15). This is due to the fact that expertise is usually restricted to a specific domain and experts have difficulties when they have to apply their knowledge to new problems. As a result, experts often fail to make better decisions under uncertainty than lay people (Chi, 2006, pp. 25–26; Lewandowsky & Thomas, 2009, pp. 150–151).
- Private interests possibly also keep peers from correctly choosing the best researcher to perform the task, meaning that the funding agency is confronted with an adverse selection problem. Influenced by their private interests, peers deliberately do not fulfil their task to reduce the information asymmetry. The problem of moral hazard on the part of the peers subsequently leads to a problem of adverse selection in the choice of the researcher. According to Travis and Collings (1991) innovative research projects are more likely subject to cognitive cronyism than mainstream research. This phenomenon leads peers to prefer supporting research that is within their area of specialization, meaning that their decisions are based on membership in school of thought (Travis & Collins, 1991, pp. 323, 336). Additionally, innovation from others is also a possible threat to the importance of the researcher's own work (Horrobin, 1990, p. 1441).
- Yet another possible source of conservatism lies in the information peers have at their disposal. If the information they possess is incomplete and does not reflect reality correctly, they cannot reduce the information asymmetry effectively even when they are willing to. Wang, Veugelers and Stephan (2016) showed that standard bibliometric measures are biased against novel research. The recognition of novel papers is delayed. When a short time window is used they are less likely to be top cited and they are usually published in journals with Impact Factors below the expectations. Thus funding decisions

based on traditional bibliometric indicators are likely to be biased against novel research (Jian Wang, Veugelers, & Stephan, 2016). The evaluation procedure of the SNSF is built on several evaluation criteria and does not solely rely on bibliometric indicators. Therefore, this possible explanation for conservatism does not hold in the context of the SNSF.

In this thesis, I assume that peers are able to identify correctly innovative proposals and describe this accordingly in their evaluations but that they fail to translate this adequately into a grade. This is either because they actively do not want innovative research to get funded because of their own private interests or because they are not able to accurately assess the opportunities of an innovative proposal (bounded rationality).
The relationship is stated separately for the external reviewers and the referees because evaluations from external reviewers are different from the ones written by referees. External reviewers only assess one application per call and therefore give a non-comparative evaluation, whereas referees by assessing more than one application per call provide a comparative evaluation.

<u>Hypothesis 2:</u> Among applications that were classified by the referees into the quality level where the funding line is drawn, the more an application is perceived as innovative, the smaller its chance of getting funding.

Langfeldt (2001) showed that the available budget and the rating scale both affect the degree to which considerations regarding distributional policy and research policy objectives (e.g. support of innovative projects) are taken into account in addition to research quality. The more funds available and the rougher the rating scale, the more likely research policies are taken into consideration. Further, the organisation of the peer review procedure also has an influence. The more thorough the process, the less likely it is that other criteria than research quality are considered. Additionally, the way decisions are reached also influences the outcome. If the members of the evaluation body are in a position achieve funding for their favourite proposals implying that one member of the evaluation body can determine some of the outcomes, chances for innovative projects are better than when a proposal is sorted out if a majority of the evaluation body does not support it (Langfeldt, 2001, pp. 830–837, 2006, pp. 36–37).

In case of the SNSF, the amounts approved in project funding grew constantly over the last ten years, from CHF 261 million in 2005 to CHF 441 million in 2015. At the same time though, the success rate dropped from 61% to 45%, indicating that it is increasingly difficult for researchers to obtain funding from the SNSF and that funding tends to be more and more reserved for highly selected projects (SNSF, 2016d, pp. 18–19).

The rating scale used by the SNSF goes from one to six. Hence, it is a rather rough scale, which leads to a situation where a lot of proposals have the same grade. This should give the evaluation body the possibility to implement policy priorities like funding for innovative projects when making the decision. Yet, it is unclear how the fact that in final meetings "the available budget is (…) distributed across the most highly-rated applications only (…)" affects the consideration of policy priorities, because particularly innovative projects might not always belong to this category (SNSF, 2016e).

The SNSF's evaluation procedure seems rather thorough and rigid, not allowing for much randomness. Decisions whether to fund or reject a proposal are made based on majority vote. If a majority is against funding a proposal it gets rejected, regardless of the support from a single evaluation body member (SNSF, 2016e).

For this reason one can suspect that overall the organization of the evaluation procedure in project funding at the SNSF might not be in favour of the promotion of innovative projects.

4 Method

There are potential approaches to determining objectively the degree of innovation of a particular project. Some authors assume that research proposals on which reviewers disagree to a great extent are likely to be innovative (Hackett & Chubin, 2003, p. 11; Kaplan, 2005). Therefore one of the possibilities to measure the degree of innovation of the applications would have been to calculate the variance of the grades attributed by the reviewers. However, Boudreau, Guinan, Lakhani and Riedl (2016), who used an experimental study design, found no evidence that novelty leads to higher variance in evaluation scores, which is why it does not seem pertinent to use this approach in this study (Boudreau et al., 2016, p. 14).

Another possible approach to determine the novelty of research proposals would be to measure whether any new combinations of referenced journals or keywords are made in the applications. However, as has been elaborated in chapter 2.3, novelty might point to innovation, but it is only one of the multiple aspects of innovation. Moreover, this approach also necessitates the gathering and treatment of enormous masses of data. Wang et al. (2016, p. 5) who assessed the novelty of journal articles, checked for every article whether its journal pairs appeared in prior Web of Science articles, starting from 1980. At the SNSF, the information regarding the articles on which the proposal builds are not accessible in a structured way.

Boudreau et al. (2016) checked the research proposals' keyword combinations against 185 million existing term combinations in PubMed. To do so they hired a professional librarian to code all the proposals according to the controlled vocabulary used in PubMed (Boudreau et al., 2016, pp. 8–9). At the SNSF, applicants are free to determine whatever keywords they want to describe their application. Hence, it would make less sense to compare keyword combinations with the ones from a database.

As these three potential approaches are neither optimal to answer my research question nor feasible in the scope of this thesis, I decided to assess the perception of the innovativeness of the research. For this purpose I analysed the reviews[7] of the applications using content analysis. Unlike the three potential approaches presented this will not produce an objective measure of the proposals' degree of innovativeness.

In the following paragraphs, the method of content analysis and the population will be described. Then the operationalization will be explained, which is followed by the description of the samples, the quality criteria and the construction of the indices.

4.1 Content analysis

Content analysis is an empirical method used to describe the content-related and formal characteristics of messages systematically and intersubjectively comprehensible (Früh, 2015, p. 29). The method is systematic because it requires drawing a sample of the population of messages applying predefined rules. Subsequently the content of these messages is to be analysed using verifiable criteria. A central element of content analysis is therefore the elaboration of a coding scheme describing the categories that should be measured. The coding scheme, usually composed of formal and content elements together with the coding instructions constitutes the operationalization of the hypotheses' variables. Formal elements are physically manifest facts, which can be collected by measuring, counting or transcribing and do not require any inference by the coder. Content elements are the structures of meaning of interest, for whose classification the coder has to make an inference.

[7] Without any further specification the term reviews always refers to the two types of reviews – the external review and the recommendation.

The definition of the categories should be based on theory. In a second step, by confronting the categories with the empirical data they can be differentiated and complemented but neither changed substantially nor reduced. Ideally the categories should be exhaustive, mutually exclusive and clearly distinctive. They should allow assessing the different aspects of a construct. Each message should clearly be attributable to one category and only relevant aspects should be captured. For the analysis to be successful it is crucial that the data model represents the structures of meaning intended in the research question (Diekmann, 2005, pp. 482, 489; Früh, 2015, pp. 32, 84, 147–157; Rössler, 2010, pp. 44, 101–102).
In order to assess evaluative statements about persons, facts or events, as is the case here with the perceived degree of innovation, categories of valuation are required. Two approaches are possible: a global assessment, where the coder has to weight and relate different aspects and circumstances himself and give a carefully considered judgement, or the assessment of single evaluative statements (Rössler, 2010, pp. 156–157, 163).

Content analysis usually focuses on one or several aspects of a communication process. The goal is to draw conclusions about the message, its producer and/or its receiver. Building on this three different approaches can be distinguished: formal-descriptive, diagnostic and prognostic. The formal-descriptive analysis deals with the formal aspects of a text only. The content is not of interest. Diagnostic analysis focuses on the relationship between the sender and the message. Its interest is on the intention of the producer of the message and the values he projects into the message. The prognostic approach attempts to investigate the effect of the message on the recipient. In order to do this additional external data has to be collected. Both the diagnostic and the prognostic approach are interpretative inferences (Früh, 2015, pp. 45–46).
In the context of this thesis, the reviews of the proposals correspond to the message. Senders of these messages are the external reviewers and the referees. The primary receivers of the messages of the external reviewers (message 1) are the referee and the co-referee, who build their own message on the ones obtained from the external reviewers (cf. Figure 1). In principle, the whole evaluation body has access to the external reviews, but it seems rather unlikely that they will read all these messages. Since around 2012, in case of negative decisions, the applicants also have access to anonymised versions of the external reviews for the purpose of transparency and information. The receivers of the messages of the referees, the recommendations (message 2) are their colleagues from the evaluation body.

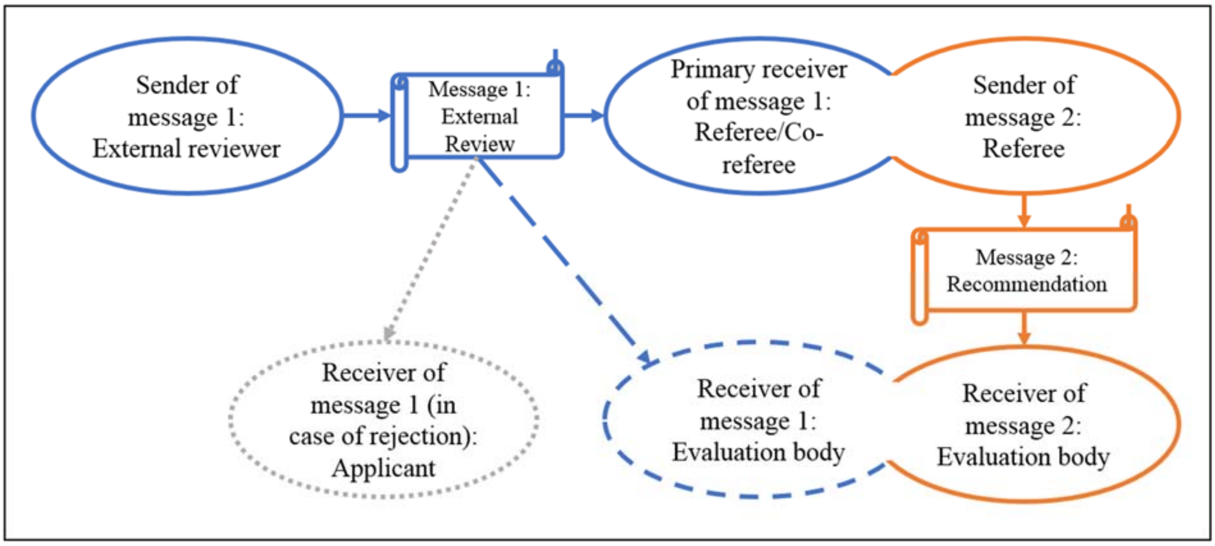

Figure 2. Communication process in the evaluation procedure (authors own figure inspired by Diekmann (2005, p. 484))

Regarding hypotheses 1a and 1b, the focus was on the messages produced. The goal was to investigate whether there is there a relationship between the degree of innovation expressed in the message and the grades given by the external reviewers and the referees. In a broad sense this corresponds to the diagnostic approach. Hypothesis 2 focused on the effect the referee's message has on the members of the evaluation body and the decision they adopt, thus, employing a prognostic approach.

4.2 Operationalization – the coding scheme

In this section the development, the different units and the elements of the final coding scheme are explained.

4.2.1 Development of the coding scheme

In a first step, existing literature regarding innovation in research was examined and a description of the concept elaborated (cf. chapter 2.3). Thereafter, guidelines for peer reviewers and descriptions of the peer review process from different funding agencies were analysed regarding their description of innovation as an evaluation criterion. They point to different aspects in which an application could be innovative e.g. concepts, approach, methods, technologies. Additionally they helped identifying terms related to innovation (AHRQ, 2014; MRC, 2016, p. 28; NHMRC, 2015, p. 2; NIH, 2016; NSERC, 2016, pp. 19, 40). Finally, the external reviews and recommendations of a subset of 20 applications were analysed in order to confront the preliminary categories with the empirical data.

After thorough consideration I decided to use a synthetic coding scheme. This specific type of coding scheme is well suited for the analysis of single evaluative statements and is best used if otherwise a large number of categories would have to be created. This was the case as one of the numerous characteristics of one content category could possibly occur together with any characteristic of the other content categories. If a category for each combination of characteristics had to be constructed, this would have amounted to more than 100 categories. With the synthetic coding scheme each characteristic of interest is analysed separately and can be assembled afterwards (Früh, 2015, pp. 214–215; Rössler, 2010, p. 163).

Every newly constructed data collection instrument should undergo a pretest in order to identify problems and adapt it so as to have a satisfying instrument (Diekmann, 2005, p. 169).
In this study 10 applications were randomly selected from the population for the pretest. All the corresponding recommendations (10) as well as all external reviews (26) were analysed. Most of the categories could be coded satisfactorily. The characteristics "novel / do something that has not been done before" and "impact" from one of the content categories were clarified and specified. In addition, the definition of what was considered a statement was further detailed. Moreover, the pretest showed that the characteristics "novel / do something that has not been done before" and "impact / advances in the academic field of study" posed a problem of overlapping. This could be resolved by further specifications in the coding instructions.

4.2.2 Definition of the different units

In order to have the content accessible for analysis an important step in the development process is to define the units (Rössler, 2010, p. 41). They are described below:

<u>Unit of data collection</u>: this defines on what level the characteristics of the variables are measured (Früh, 2015, p. 91). In the context of this thesis the application itself was the first unit of data collection. The second unit was the external review, respectively the recommendation. The content elements were coded on the statement level. Each statement on innovation in a review was coded according to the content characteristics described in the coding scheme.
<u>Unit of analysis:</u> the external review, respectively the recommendation was the element on which data was analysed and for which findings were reported (Früh, 2015, p. 91). Afterwards the number of coded statements was calculated and weighted so that it reflected the degree of innovation as perceived by the reviewer (cf. chapter 4.6 for a description of the construction of the indices).

Unit of sampling: this systematically defines what material is to be analysed content-analytically. Statistically speaking it is the sample that has been drawn from the population. For this thesis, two random stratified samples were used (cf. section 4.4) (Rössler, 2010, pp. 42, 61).

4.2.3 Coding scheme: formal elements

The formal elements are those that serve as control variables. They regarded the application itself, the responsible applicant, the review by the external reviewer and the recommendation by the referee as well as the appearance of the reviews (cf. Annex 1 for the coding scheme used).

Concerning the application the following data was collected:

- The *funding status* of the application. Was the application approved or rejected?
- The *year* the letter of ruling was sent out. Success rates usually differ slightly from one year to another (SNSF, 2016d, p. 18).
- The *discipline group* to which the applicant attributed his application. There are seven discipline groups in biology and medicine, each consisting of a multitude of disciplines: basic biological research, general biology, basic medical sciences, experimental medicine, clinical medicine, preventive medicine and social medicine. This information is relevant because the success rate varies across the groups of disciplines (SNSF, 2016d, p. 12).
- The *type of institution* at which the project was to be carried out, as the probability of funding differs significantly across the four main institution types (ETH domain, cantonal universities, universities of applied sciences and others). The probability of being funded is highest for applications from the ETH domain, followed by cantonal universities and others. Applications of universities of applied sciences have the smallest chances of getting funding (Coryn et al., 2012, p. 28).

With reference to the responsible applicant the following categories were considered relevant:

- The applicants' *gender*, because gender might have an influence on the success rates. The evaluation by Coryn et al. (2012, p. 27) found that the chances of getting funded for male applicants are slightly higher than for female applicants[8]. It is also one of the variables often associated with bias in peer review.

- The *applicant's age* at the time of submission, as the success rate might be different across the age categories. Established researchers are given an advantage over younger researchers. Merton (1968) described this as the "Matthew-Effect in Science".
- The *academic degree*, as the "Matthew-Effect" also applies to this case. Applicants holding a professorship usually already have funding and therefore might even receive more compared to applicants holding a PhD (Merton, 1968).

As for the review by the external reviewer the following data was collected:

- The *country of residence* of the external reviewer is of importance because reviewers from abroad tend to give higher grades than reviewers based in Switzerland (SNSF, 2016a, p. 7).
- The *source of recommendation*, because reviewers recommended by the applicant also give better grades compared to reviewers chosen by the SNSF (either after recommendation by the referee, the Administrative Offices or another reviewer, SNSF, 2015a, p. 7)[9].
- The *expertise of the reviewer* as declared by the reviewer himself. The reviewers state whether the application falls into their area of specialisation or whether it is within their wider discipline. Two studies found that proposals were rated lower by reviewers with a shorter intellectual distance from the proposal (Boudreau et al., 2016, p. 2; Gallo, Sullivan, & Glisson, 2016, p. 14).
- The *usefulness* of the review as judged by the referee, because only reviews deemed useful are going to be analysed.
- The *number of external reviewers* who assessed an application.
- The *grades* given for the applicant's scientific track record and expertise; the scientific relevance, originality and topicality; the suitability of methods and feasibility and the overall assessment.

[8] "The SNSF analyses the differences between the success rates of female and male applicants every year" (SNSF, 2015b, p. 33).

[9] Since 01.01.2016 the applicants no longer have the possibility to recommend potential reviewers for their own applications (cf. art. 25 par. 5 of the Funding Regulations of 27.05.2015 in comparison to art. 18 par. 7b of the Funding Regulations of 14.12.2007).

Regarding the recommendations by the referees the following data was collected:

- The *grades* given for the scientific track record and expertise of the applicants, the assessment of the proposed project (with reference to the SNSF criteria) and the comparative ranking.

The appearance of the reviews was also assessed via two variables:

- The *length* of the external review / recommendation.
- The *quality* of the external review / recommendation.

4.2.4 Coding scheme: content elements

As has been noted above, the content was assessed using a synthetic coding scheme. It consists of four categories, of which each was coded once for every statement identified:

- The *location* of the statement: in what part (corresponding to the blocks of criteria) of the external review / recommendation was the statement made. As the goal was to measure an application's degree of innovation it was renounced to code the text in the criteria block about the applicant's scientific track record and expertise. This implies that the second kind of arguments, the character of the applicant, which according to Lamont et al. (2007, p. 177) is invoked in the assessment of the originality of research, was not considered (cf. chapter 2.2.3).
- The *term* used to describe innovation: what term did they use to describe innovation. A list of synonyms describing aspects of the application itself, accompanied by a term referring to the emotions of the reviewers regarding the application, i.e. the first and the third kind of arguments as described by Lamont et al. (2007, p. 177) were included. As a third type, terms referring to the potential impact of the application were also included. This is of importance, as an innovation according to its definition also requires adding value.
- The *valuation* of the statement: how did the reviewers value their statement about innovation. Four different values were distinguished: very positive, positive, negative and very negative. As in grant peer review most of the comments usually are positive (cf. chapter 2.2.3) already statements saying that something is not very innovative or where doubts about the innovativeness were mentioned were judged as being negative.

- The *aspect of the application* that is described as being innovative. Does the statement refer to the project in general or to a clearly identifiable aspect (e.g. the topic or its approach)?

4.3 Population

The population in the context of this study consists of the reviews received for the applications submitted to the division biology and medicine for project funding from 2012 to 2016. During this time period the evaluation procedure and the criteria to be applied stayed the same. Applications prior to 2012 have not been considered for the reason of the comparability because the improved guidelines for the assessment by external reviewers were published in late 2011. Because only applications for which the evaluation procedure has been finished are relevant, applications from the second call in 2016 are not taken into account. Furthermore, with the second call in 2016, project funding underwent major changes and those cases would not necessarily be comparable to previous years. The analysis was restricted to one division in order to only have applications that were reviewed by one division of the NRC. The three divisions of the NRC most likely differ in their function and as Langfeldt (2001) showed that this influences to what extent research policy objectives are considered. The division biology and medicine has been selected because Reinhart's study (2012) also used data from the said division and because it seems plausible that a considerable part of the applications in this division is innovative.

Applications for which the evaluation procedure is likely to differ from the standard process and applications that did not undergo the evaluation procedure have been excluded from the population[10]. A total of 922 applications fulfilled these selection criteria, of which 428 have been accepted[11]. For the 922 applications, a total of around 2'600 reviews were written. This corresponds to an average of 2.9 reviews per application, which is very close to the SNSF-wide average of 3.0 reviews per application (SNSF, 2016a, p. 2).

[10] Specifically, the following types of applications have been excluded: Follow-up applications, resubmissions, lead agency applications, bonuses of excellence, directly rejected applications, applications that were withdrawn by the applicant before the evaluation and applications for which a formal decision not to examine the content had been taken.

[11] It should be noted that the corresponding success rates are not directly comparable to the success rates published in the annual reports due to the differing data selection criteria.

4.4 Samples

In order to test the two sets of hypotheses, two different random stratified samples have been drawn. The random stratified sampling method is well suited when the distribution of a variable of interest in the population is known to be uneven. It assures that all the characteristics of a variable of interest are present in the unit of sampling, even if only a limited number of cases is analysed. If the sample size of the strata corresponds to their relative proportion in the population it is called a proportionate stratified sample otherwise it is denominated disproportionate. In the case of a disproportionate stratified sample, an appropriate weighting has to be undertaken to make a statistical inference for the population (Früh, 2015, pp. 102–103; Neuendorf, 2002, pp. 85–86).

Sample 1 (hypotheses 1a and 1b)

In a preliminary analysis the type of institution was identified as the variable being the most interesting and relevant. An evaluation study performed in 2012 found that the success rate differs significantly among the different types of institutions (Coryn et al., 2012, p. 28). Hence, four strata, one for each type of institution, were created. From each stratum 30 applications were selected randomly. This resulted in variable sampling ratios, meaning that the probability of an application to be included in the sample differed among the different strata (cf. Table 1). The sample is disproportionate (Maletta, 2007, p. 4). Considering the universities of applied sciences, only 24 applications existed and all were included in the sample. Their probability of being selected was 100%, whereas, for example, the probability of applications from universities being selected was much lower, as there were a total of 701 applications of which only 30 were selected for this analysis. The total sample size amounted to 113[12]. The corresponding 113 recommendations (later referred to as sample 1: recommendations) plus one randomly selected external review per application (later referred to as sample 1: external reviews) were analysed[13].

[12] For one of the applications from the universities of applied sciences I could not access the recommendation or the external reviews. Which is why only 23 of the 24 applications from the universities of applied sciences were included effectively in the sample.

[13] In the case of 16 applications in the corresponding population the referee was not a member of the NRC but from a panel. As all of these applications came from universities of applied sciences, they were nevertheless all included in the sample. This was done in order not to have an even smaller sample size.

Table 1. Stratification of sample 1

	N	%	n	%	Sampling Ratio	Sampling Weight	n_w	$\%_w$
ETH domain	138	15.0	30	26.5	0.2174	0.5638	17	15.0
Cantonal universities	704	76.4	30	26.5	0.0426	2.8638	86	76.0
UAS	24	2.6	23	20.4	0.9583	0.1279	3	2.6
Others	56	6.1	30	26.5	0.5357	0.2410	7	6.4
Total	922	100.0	113	100.0			113	100.0

Sample 2 (hypothesis 2)

In the division biology and medicine applications are usually funded only up to the third quality level. The funding line is drawn among applications rated with a B: 50% of the proposals are weaker, 25% stronger. For this reason only applications that were attributed to the third quality level were considered (277). Among this subset, two strata were created, one with the funded (200) and one with the rejected applications (77). Again 30 applications were selected randomly from each stratum, leading to a disproportionate sample with differing sampling ratios (cf. Table 2)[14]. In this case only the corresponding recommendations were analysed.

Table 2. Stratification of sample 2

	N	%	n	%	Sampling Ratio	Sampling Weight	n_w	$\%_w$
Approved	237	73.4	30	50.0	0.1266	1.4675	44	73.4
Rejected	86	26.6	30	50.0	0.3488	0.5325	16	26.6
Total	323	100.0	60	100.0			60	100.0

4.5 Quality criteria in content analysis

In content analysis the most important quality criteria are reliability and validity. Reliability describes the extent to which a measuring procedure generates the same results on repeated trials. Validity describes the degree to which a measuring procedure represents what it intended to measure. The two criteria are in a hierarchical relationship. Reliability is a necessary but not a sufficient condition for validity, whereas validity is not a necessary condition for reliability (Diekmann, 2005, p. 227; Rössler, 2010, p. 116). In the following, the two criteria will be further described and it will be explained how this thesis met them.

[14] In the case of 4 applications in the corresponding population the referee was not a member of the NRC but from a panel. None of them were selected for analyses.

4.5.1 Reliability

As content analysis is defined as an intersubjectively comprehensible method, measuring the reliability is a standard. There are three types of reliability that can be distinguished: intercoder reliability, intracoder reliability and researcher-coder reliability. In the context of this thesis only intracoder reliability was relevant, because there was only one coder, who at the same time was the researcher. It measures how well the coding at the beginning of the data collection matches the coding towards the end of the data collection. To assess this, the coder has to analyse some material he analysed in the beginning again at the end of the data collection (Rössler, 2010, pp. 197–198).

In order to test the reliability of the data coded, a subset of 25 randomly selected recommendations respectively external reviews from the two samples was assessed a second time, about two weeks after the data collection. The time between the two data collections should ensure that the coder could not remember the coding made when the content was assessed for the first time. By selecting 25 reviews of each type, it was assured that for the content categories, which were coded on the statement level, the generally required minimum of 30 to 50 codes per category could be obtained. For the formal categories, which were coded on the level of the application, logically only 25 codes were assigned per type of review. But as these categories do not require inference this was not problematic.

Since the coder had to identify the unit of analysis, a two-stage reliability analysis was applied for the content categories. In a first step the identification reliability was calculated. It indicates the correspondence in the identification of the statements to be coded. This is important, as the material to be coded has to be selected in a reliable manner. Secondly, the coder reliability was computed for the matching identified units. Only the exact same code was considered a match. For the calculation the widely used Holsti reliability coefficient was applied. It relates the number of matching codes to the total number of codes and gives back values from 0 to 1 (Früh, 2015, pp. 181–187; Rössler, 2010, pp. 198–205). In general, reliability was sufficiently high for all the variables used for the analysis[15]. The identification reliability was also rather high with values of $C_r = 0.987$ for recommendations and $C_r = 0.965$ for external reviews. With $C_r = 0.908$ the reliability coefficient was lowest, but still satisfying for the variable "Aspect of the application described as being innovative" (cf. Annex 2).

[15] It was also attempted to assess the perceived degree of innovation globally, as a summary judgement, which is known to be a complex issue (Rössler, 2010, p. 157). The reliability coefficients were below 0.70 and therefore unsatisfactory, which is why finally the use of this variable was refrained from.

4.5.2 Validity

Validity indicates whether and to what extent a theoretical construct has been well operationalized. More precisely, it answers the question of whether a coding scheme measures what the research question intended. Diekmann (2005) describes three forms of validity: content validity, criterion-related validity and construct validity. Content validity is present if the measure reflects the relevant aspects of the concept being measured. Criterion-related validity describes to what extent the result of a measuring procedure correlates with other relevant characteristics. These characteristics need to be obtained independently through a distinct measuring procedure. Construct validity shows in what way a measuring procedure is useful to generate a theory. It demands that a measured construct is related to as many possible other variables in a way consistent with a theory (Diekmann, 2005, p. 224).

Content validity was ensured with the confrontation of the data material with the coding scheme during its elaboration and the pretest, which both allowed for the integration of additional categories and the fact that an *other* residual category was used throughout the data collection. The assessment of criterion-related validity is relatively difficult and a statement its respect cannot be made. Construct validity is supported by the fact that the coding scheme was developed based on theory and existent peer review guidelines.

4.6 Construction of indices

Using the coded content elements, two main indices were built and used throughout the study as the variables of interest: „perceived degree of innovation“ and „anticipated impact“.

For the index “perceived degree of innovation”, the valuations of the statements including a synonym for innovation were summed up per identified aspect of the application (project in general, research goal / topic, theory and approach / method). In order not to overestimate the degree of innovation, especially in longer external evaluation / recommendation, the mean was calculated per aspect. In the last step these means were added up again to form a total measure of the perceived degree of innovation. The possible maximum and minimum values amount to 4 respectively -4, which would mean that all the aspects are valued very positive or very negative[16].

[16] Example: an external review consists of three statements: a very positive statement (value=2) using the term originality and mentioning the project in general; a positive statement (value=1) using the term innovative mentioning the project in general and a positive statement (value=1) using the term original and mentioning the approach. In this example the perceived degree of innovation would amount to 2.5 ((2+1)/2 + 1/1).

The index "anticipated impact" was constructed in a similar way. First, the valuations of the statements referring to the potential impact of the application were added up per different type of impact mentioned (impact on patient / treatment, impact / advances in the academic field and other impact (which consisted of remarks about the (broader) impact in general and statements about the impact on policy). Then again, the means were calculated for each type of impact, which in the end were added up to build the index "anticipated impact". If all three types of impact were valued very positive or very negative, the possible maximum and minimum values would add up to 3 respectively -3.

By calculating the indices in the manner described, it is possible that the value obtained is zero even though the reviewer made statements about the proposals innovativeness or its potential impact (e.g. when the mean for the project in general is 1 and the mean for the approach / method is -1). This case did not occur very often. Therefore it was treated as if the value was zero because there was no statement at all. The advantage of calculating the indices in this manner is that it prevents overestimating the degree of perceived innovation, as repeated or very similar statements do not count as much as they would were the statements simply added up.

5 Descriptive analysis

In this chapter, the composition of the samples is analysed first and compared to the population basis. Then the characteristics of reviews will be presented, followed by an analysis of the innovation related statements identified in the reviews and the values of the indices “perceived degree of innovation” and “anticipated impact”.

5.1 Comparison of the samples with the population

Sample 1

Of the 113 applications in the sample, 45% were approved[17]. This corresponds to the success rate observed in the population. Regarding the final ratings by the NRC, applications rated as AB were underrepresented in the sample with only 9% (population 17%), whereas those with a BC rating were overrepresented (32% vs. 24% in the population). The discipline groups, the applicant’s gender, the academic degree and the age category were well represented in the sample. The biggest difference between the sample and the population presented was the percentage of the applicants in the age category of 41-45 years old (37% vs. 27% in the population) (cf. Annex 3 Tables 5.1-5.6).

Sample 2

Of the 60 applications in the sample, which all were rated a B by the referees, 73% were approved. By means of weighting, this corresponded to the success rate observed in the respective population (cf. Table 2 above). With regard to the institutional affiliation, applications from the ETH domain were overrepresented (29% vs. 19% in the population) and applications from cantonal universities slightly underrepresented (65% vs. 71% in the population). In this sample the discipline groups, the applicant’s gender, the academic degree and the age category were well represented too. The biggest difference between sample and population could be found in the applicant’s academic degree. Applicants with doctoral degree were underrepresented (38% vs. 51% in the population), whereas professors were overrepresented (62% vs. 49% in the population) (cf. Annex 3 Tables 5.7-5.11).

[17] All the numbers and results reported in this and the following chapters were weighted by the sampling weights reported in Table 1 and Table 2. The only exception represents numbers regarding the institution type (in case of sample 1) or the funding status (in case of sample 2). In these cases weighting is not necessary as the stratified samples were based on these variables. In the tables weighted values are marked with subscripted lower case w (e.g. n_w).

5.2 Characteristics of the reviews

Sample 1: External reviews

The majority of the external reviews were between 251-1'000 words long. They were comparable in length to the reviews from 1998 analysed by Reinhart (2012, p. 160). The coder judged around 86% of the reviews as being of good quality. Even though some were judged as poor quality they were still included in the analysis (cf. Annex 3 Tables 5.12-5.13).

Most of the external reviewers were chosen by the SNSF (87%, either by recommendation by the Administrative Offices, the referee or another reviewer). In 13% of cases they were on the applicant's list (cf. Annex 3 Table 5.14). Altogether the distribution among the sources of recommendation corresponds with the distribution observed at the SNSF as a whole, as was reported in the internal Monitoring Reports on Peer Review (SNSF, 2016a, p. 4).

The referees judged the large part of the external reviews in the sample as being useful (84%). They only judged 16% as only being useful in part (cf. Annex 3 Table 5.15).

On average, for each proposal 2.5 reviews were obtained. This is somewhat below the average of 2.9 reviews per proposal in the population (cf. Annex 3 Table 5.16).

Researchers residing outside of Switzerland provided most of the external reviews in the sample (63%). Still, reviewers working in Switzerland provided more than one third (37%) of the external reviews (cf. Annex 3 Table 5.17). At the SNSF in general, the share of Swiss-affiliated reviewers is lower (SNSF, 2016a, p. 7).

Regarding their expertise, 59% of external reviewers indicated that the proposal is within their area of specialisation (i.e. they have high expertise). The remaining 41% reported that the proposal is within their wider discipline (i.e. they have lower expertise, cf. Annex 3 Table 5.18).

Sample 1: Recommendations

The majority of the recommendations made were between 100-750 words long. The coder defined almost all the recommendations as being of good quality (99%) (cf. Annex 3 Table 5.19-5.20).

On average, it could be observed that the external reviews were longer than the recommendations by the referees (666 vs. 384 words) (cf. Annex 3 Table 5.21). However, it should be noted that in case of recommendations only the text in the criteria-based assessment was counted, not taking into account the summary of the project outline and the external reviews, which makes up for a considerable part of the recommendation's text.

Sample 2

The recommendations in the second sample were very similar in length and quality as the recommendations from the first sample. The majority of the recommendations were also between 100-750 words long, but slightly longer on average (410 vs. 384 words). Regarding the quality, 93% of these recommendations were judged as being good (cf. Annex 3 Tables 5.21-5.23).

5.3 Analysis of the identified statements

In this section the statements identified as being related to innovation in the reviews will be analysed descriptively. Their occurrence per review, their distribution among the different parts of the reviews, the terms used, the valuation and the aspect described as being innovative will be presented.

Sample 1: External reviews

In total 316 statements[18] were identified in the text of the 113 external reviews analysed. In the case of 27% of external reviews no statement about innovation could be found. For 62% of the external reviews one to five statements were identified. Only 11% contained more than five statements (cf. Annex 3 Table 5.25). This is similar to the results Reinhart (2012) obtained. He found that 66% of the analysed reviews contained statements about the application's originality (Reinhart, 2012, p. 168).

Most of the statements were identified in the text box dealing with the scientific relevance, originality and topicality (63%). This was followed by the comments regarding the overall assessment, which contained 28% of the statements. Just 9% of the statements were located in the section about the suitability of methods and feasibility (cf. Annex 3 Table 5.26).

The term most often identified in a statement was 'original' (28%), followed by statements about the impact the proposal would have in the academic field (26%). The third most frequent term was 'novel' (18%). Other terms were identified much less often (cf. Table 3).

[18] Due to the weighting, the number of statements reported differ from the number of statements originally observed (cf. Annex 3 Tables 5.24).

Table 3. Statements by terms used (Sample 1: External Reviews)

	n_w	$\%_w$
Innovative	17	5.5
Original	89	28.2
Novel	55	17.5
Creative	8	2.6
Groundbreaking	0	0.0
Impact in general	7	2.1
Impact on patient / treatment	24	7.6
Useful for policy-/ decision makers	3	1.1
Impact / advances in the academic field	83	26.3
Exciting	11	3.6
State of the art	5	1.6
Cutting edge	4	1.4
Risky	8	2.4
Other	1	0.2

Regarding the valuation, positive statements were by far the most frequent (78%), followed by very positive statements (13%). Only 9% of the statements were negative or very negative (cf. Annex 3 Tables 5.27). This corresponds relatively well to the finding of Reinhart, who found that 67% of statements about originality were positive, 27% negative and 6% neutral (Reinhart, 2012, p. 168).

A total of 59% of the statements were about the project in general. In 30% of the cases, the approach / method was described. Further, 10% of the statements were about the research aims / topic, and one statement was not attributable to any of the aspects. No statement referred to the theory / hypotheses (cf. Annex 3 Table 5.28).

Sample 1: Recommendations

In total, 139 statements were identified in the text of the 113 recommendations in the sample. In the case of 41% of recommendations no statement about innovation could be found. For the largest part of the recommendations (45%), one to two statements were identified. Only 14% contained more than two statements (cf. Annex 3 Table 5.29).

In the text box about the assessment of the project more statements were identified than in the comments regarding the comparative ranking (65% vs. 35%) (cf. Annex 3 Table 5.30).

The term most often identified in a statement was also ‘original’ (31%), followed by statements about the impact the proposal would have in the academic field (23%). The third most frequent term was ‘novel’ as in the external reviews (15%). All the other terms were identified less often (cf. Table 4).

Table 4. Statements by terms used (Sample 1: Recommendations)

	n_w	$\%_w$
Innovative	11	7.7
Original	43	31.2
Novel	20	14.8
Creative	0	0.0
Groundbreaking	0	0.1
Impact in general	2	1.3
Impact on patient / treatment	10	7.2
Useful for policy-/ decision makers	3	2.5
Impact / advances in the academic field	32	23.2
Exciting	2	1.6
State of the art	2	1.3
Cutting edge	7	5.1
Risky	6	4.2
Other	0	0.0

Regarding the valuation, positive statements were also the most frequent (60%). But contrary to the case of the external reviews followed by negative statements (27%). In 6% of the statements the comments were either very negative or very positive (cf. Annex 3 Table 5.31). Further, 79% of statements were about the project in general. This was followed by 14% of statements about the approach / method and 7% of statements about the research aims / topic. There were no statements about the theory / hypotheses and only a negligible part was not attributable to a precise aspect (cf. Annex 3 Table 5.32).

The recommendations less often contained statements referring to innovation (59% of analysed recommendations) than external reviews (73%). On average, they also contained fewer statements than external reviews (1.2 vs. 2.8). However, recommendations seem to be more critical than external reviews as they more often contain negative statements (26% vs. 16% of analysed reviews) (cf. Annex 3 Tables 5.29, 5.25, 5.24, 5.31 and 5.27).

Sample 2[19]

In total, 91 statements were identified in the text of the 60 recommendations analysed for the second sample. In 26% of these recommendations no statements about innovation could be found. For 55% of the recommendations one to two statements were identified. Only 9%

[19] The content of recommendations from sample 2 differed from the content of recommendations from sample 1, as all of them indicated that the applications belongs to the third quality level, whereas the applications associated with the recommendations from sample 1, were ranked as belonging to different quality levels.

contained more than two statements. The average number of statements amounted to 1.5 (cf. Annex 3 Table 5.33 and 5.24).

About the same number of statements were identified in the text box about the assessment of the proposed as in the comments regarding the comparative ranking (51% vs. 49%) (cf. Annex 3 Table 5.34).

The term most often identified in a statement was again ‘original’ (43%), followed by statements including the term novel (21%). The third most frequent term was ‘innovative’ (9%). The other terms were identified less often (cf. Table 5).

Table 5. Statements by terms used (Sample 2)

	n_w	$\%_w$
Innovative	9	9.3
Original	39	43.2
Novel	19	21.0
Creative	3	3.2
Groundbreaking	3	3.8
Impact in general	4	3.9
Impact on patient / treatment	3	2.8
Useful for policy-/ decision makers	0	0.0
Impact / advances in the academic field	6	6.3
Exciting	0	0.0
State of the art	3	3.8
Cutting edge	0	0.0
Risky	3	2.8
Other	0	0.0

Regarding the valuation, positive statements were the most frequent (66%) again followed by negative statements (22%). Very positive were 9% of the statements and 2% very negative (cf. Annex 3 Table 5.35).

A total of 41% of the statements were about the project in general, followed by 29% of the statements being about the approach / method and 28% of the statements were about the research aims / topic. Only 2% of the statements were about the theory / hypotheses (cf. Annex 3 Table 5.36).

5.4 Analysis of the indices “perceived degree of innovation” and “anticipated impact”

Sample 1: External reviews

The index “perceived degree of innovation” was negative in 9 of the 113 cases (8%). In 30% of the cases it was zero, either because there were no statements or because the calculation produced a zero (cf. remark in chapter 4.6). The value was between zero and two in 28% of the cases. For most of the external reviews (35%) a value of two or greater was obtained (cf. Table 6). On average the index “perceived degree of innovation” amounted to $\overline{x}_{\text{innov}}$ = 0.935 (cf. Annex 3 Table 5.37).

Table 6. Frequency distribution of the indices (Sample 1: External reviews)

	Index “perceived degree of innovation”		Index “anticipated impact”	
	n_w	$\%_w$	n_w	$\%_w$
Value is negative	9	7.8	4	3.4
Value is 0	33	29.6	50	44.6
Value is between 0 and 2	32	28.1	40	35.4
Value equals 2 or greater	39	34.5	19	16.6

The index “anticipated impact” was less often negative (3%) and zero more often (45%) than the index “perceived degree of innovation”. In 35% of the cases its value was between zero and two. Only in 17% of the cases was it equal to two or greater (cf. Table 6). Therefore, its mean was also somewhat lower with $\overline{x}_{\text{impact}}$ = 0.763 (cf. Annex 3 Table 5.37 and Tables 5.38-5.39 for statistics about the components of the indices).

Sample 1: Recommendations

The index “perceived degree of innovation” was negative in 22 of the 113 recommendations (19%). For the largest part it was zero (55%). The value was between zero and two in 24% of cases and was equal or greater than two in only 2% of cases (cf. Table 7). This resulted in a mean of only $\overline{x}_{\text{innov}}$ = 0.017 (cf. Annex 3 Table 5.40).

The index “anticipated impact” was negative less often (3%), but it was zero even more often (79%) than the index “perceived degree of innovation”. In 10% of the cases their value was between zero and two. For 8% of the recommendations it was equal or greater than two (cf. Table 7). Its mean was somewhat higher with $\overline{x}_{\text{impact}}$ = 0.233 (cf. Annex 3 Table 5.40 and Tables 5.41-5.42 for statistics about the components of the indices).

Table 7. Frequency distribution of the indices (Sample 1: Recommendations)

	Index "perceived degree of innovation"		Index "anticipated impact"	
	n_w	$\%_w$	n_w	$\%_w$
Value is negative	22	19.2	4	3.1
Value is 0	62	55.1	89	78.8
Value is between 0 and 2	27	23.5	11	9.6
Value equals 2 or greater	3	2.3	10	8.4

Sample 2

In 8 of the 60 recommendations from the second sample (13%) the index "perceived degree of innovation" was negative. In 32% of the cases it was zero. The value was most often between zero and two (37%) and was equal or greater than two in 18% of cases (cf. Table 8). This resulted in a mean of $\overline{x}_{\text{innov}} = 0.690$ (cf. Annex 3 Table 5.43).

Again the index "anticipated impact" was less often negative (4%), but it was zero in the majority of the recommendations (83%). In 13% of the cases its value was between zero and two. There were no values equal or greater than two (cf. Table 8). With $\overline{x}_{\text{impact}} = 0.073$ its mean was lower than the mean of the index "perceived degree of innovation" (cf. Annex 3 Table 5.43 and Tables 5.44-5.45 for statistics about the components of the indices).

Table 8. Frequency distribution of the indices (Sample 2)

	Index "perceived degree of innovation"		Index "anticipated impact"	
	n_w	$\%_w$	n_w	$\%_w$
Value is negative	8	12.9	3	4.4
Value is 0	19	31.8	50	82.9
Value is between 0 and 2	22	37.1	8	12.7
Value equals 2 or greater	11	18.2	0	0.0

6 Inferential statistical analysis

In this chapter the statistical analyses that were performed to test the hypotheses are presented. The models used, the assumptions for analyses and how the data met them will be explained. Thereafter, the results will be presented and discussed in the light of the postulated hypotheses.

6.1 Influence of the expressed perception of innovation on the grades

The relationships postulated in hypothesis 1a and 1b were tested using ordinal logistic regression analyses. This method allows for the prediction of an ordinal dependent variable like the grades by the external reviewers or the referees, based on one or many independent variables. By using this procedure one can detect which independent variables have a statistically significant effect on the dependent variable and determine how well the model predicts the dependent variable (Lærd Statistics, 2015b).

6.1.1 Description of the models used to test H1a and H1b

Hypothesis 1a

Figure 3 illustrates the model that was used to test hypothesis 1a (Model 1a). It postulates the following relationship: the more a proposal is perceived as innovative by an external reviewer, the lower the grade given by the reviewer.

It was anticipated that the indices "perceived degree of innovation" and "anticipated impact", which were used as independent variables are negatively associated with the grade given for the overall assessment, which was used as the dependent variable.

In order not to overestimate or underestimate the effect of the independent variables, the applicant's characteristics, the information related to the application and the features regarding the external reviewer were entered as control variables. These control variables were chosen because they have been shown to be of influence in previous analyses. See chapter 4.3.3 for a description of the control variables and the justification for their inclusion.

In a second step the model was modified in that it included the single components of the indices instead of the indices themselves; the aspects of the application; project in general, research goal / topic, theory, approach / method and the different types of impact; the impact on patient / treatment, the impact / advances in the academic field, other impact. This will be referred to as Model 1.1a.

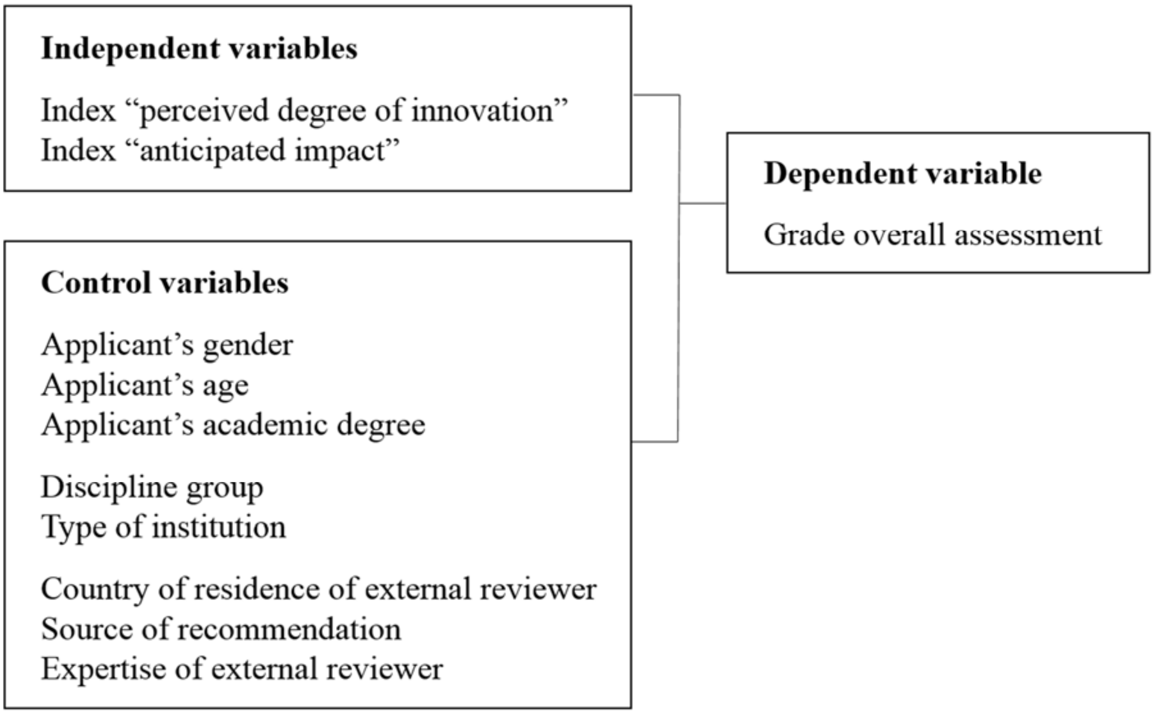

Figure 3. Illustration of Model 1a

Statistical hypothesis

H_0: $OR_{innov} \geq 1$ and $OR_{impact} \geq 1$ **H_1:** $OR_{innov} < 1$ and $OR_{impact} < 1$

Hypothesis 1b

The model that was used to test hypothesis 1b is illustrated in Figure 4 (Model 1b). This hypothesis postulates the following relationship: the more a proposal is perceived as innovative by a referee, the lower the grade given by the referee.

As in model 1a it was anticipated that the indices “perceived degree of innovation” and “anticipated impact” are negatively associated with the grade given by the referee for the comparative ranking, which was used as the dependent variable.

The applicant’s characteristics and the information related to the application were entered as control variables in order to correctly estimate the effect of the independent variables.

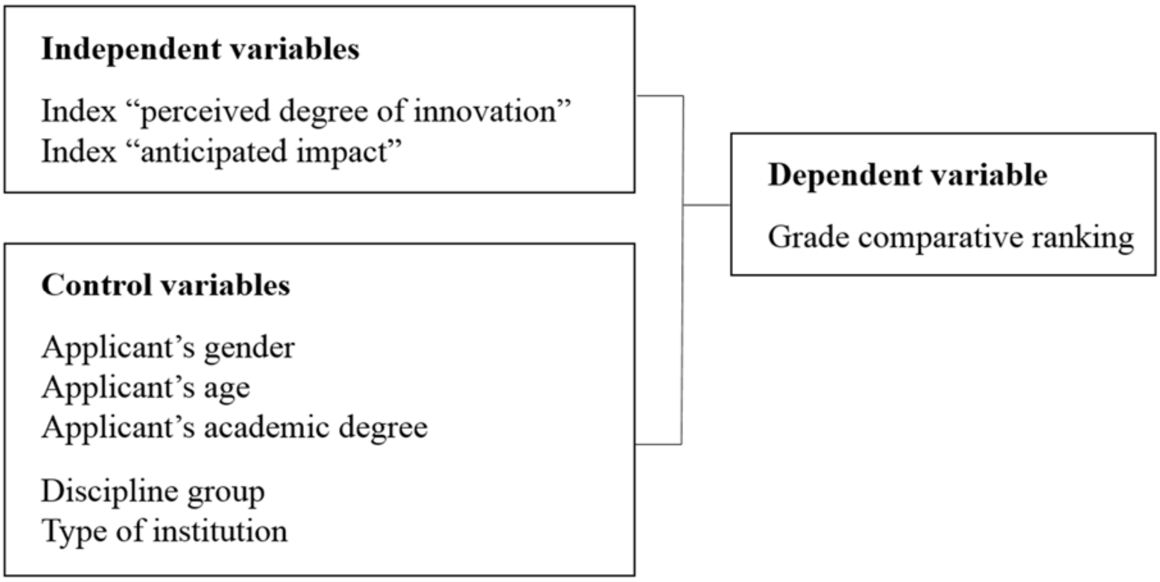

Figure 4. Illustration of Model 1b

Statistical hypothesis

H_0: $OR_{innov} \geq 1$ and $OR_{impact} \geq 1$ **H_1:** $OR_{innov} < 1$ and $OR_{impact} < 1$

6.1.2 Requirements of an ordinal logistic regression

Some requirements need to be met to perform an ordinal logistic regression analysis and obtain unbiased, efficient estimates. The most important assumptions are listed below (Lærd Statistics, 2015b). It is outlined how the existing data met them and if necessary what further steps were taken.

1) *Dependent variable measured at the ordinal level.*
   The grade given by the external reviewer / referee constituted the dependent variable. It was measured at the ordinal level.
2) *Independent variables measured on a continuous, ordinal or categorical scale. Ordinal variables must be treated as if they were continuous or categorical.*
   The variables included in the models met this requirement.
3) *No multicollinearity.*
   The independent variables should not be highly correlated with each other. In order to test for multicollinearity, the Variance Inflation Factor (VIF) can be calculated. If it is smaller than 10, it can be assumed that there is no multicollinearity. The VIF-values were smaller than 10 for all the models estimated (cf. Annex 4 Tables 6.1-6.3).

4) Proportional odds.

The independent variables should have an identical effect at each cumulative split of the ordinal dependent variable. This assumption can be tested in SPSS Statistics with a full likelihood ratio test that compares the fit of the model with proportional odds to a variable location parameters model. The two models estimated to test hypothesis 1a met this requirement. The test of parallel lines was not statistically significant (cf. Annex 4 Tables 6.4 and 6.5).

This assumption posed problems for Model 1b. For the whole model the test of parallel lines was statistically significant (cf. Annex 4 Table 6.6). As recommended in this case, I ran a multinomial logistic regression. Which has the drawback that the ordinal nature of the dependent variable is lost. As the results of the multinomial approach turned out similar to those using the ordinal approach and because the test of parallel lines was not statistically significant when only the two indices were entered, I decided to continue with reporting the results of the ordinal logistic regression analysis. This also allows a greater comparability of the results of the models used to test H1a and H1b.

I refrained from expanding Model 1b to include the components of the indices due to the problem with proportional odds and because there were even more cases with zero values than in the case of the external reviews.

6.1.3 Results from the ordinal logistic regression analyses

The ordinal logistic regression analyses were calculated using the PLUM procedure in IBM SPSS Statistics 21. In this section different methods to assess the model fit and their results will be presented separately for the two hypotheses. This is followed by the presentation of the parameter estimates and their interpretation.

Hypothesis 1a

For both models more than 80% of cells showed zero frequencies. This makes the use of a chi-squared test problematic. Therefore, the goodness-of-fit tests are not reported.

The Nagelkerke coefficient of determination indicates the percentage of the variance explained through the ordinal logistic regression. Model 1a explained 78% of the variance and Model 1.1a 82% (cf. Tables 9 and 10).

The likelihood-ratio test, which looks at the change in the model fit when comparing the full model to the intercept-only model was significant for both models (cf. Tables 9 and 10). The variables entered add statistically significantly to the model. In other words, at least one of them

is statistically significant (Bühl, 2010, pp. 442–443).
Another way to test whether the model fits the data well is to see how well it predicts the correct category of the ordinal dependent variable. For this purpose a confusion table based on the observed and predicted categories was calculated (Lærd Statistics, 2015b). Model 1a correctly predicted 71% of the cases. This value rose to 83% when the components of the indices were entered as independent variables in Model 1.1a (cf. Annex 4 Tables 6.7 and 6.8).

The results of the overall omnibus statistical tests showed that between the two variables of interest in Model 1a, only the index "anticipated impact" had a statistically significant effect on the prediction of the grade given by the external reviewers. The „perceived degree of innovation" had no significant influence. An increase in the "anticipated impact" however, was associated with an increase in the odds of getting a higher grade (cf. Table 9). Based on this result, hypothesis 1a is rejected.

Table 9. Model 1a: extract of results

	Sig.	Odds Ratio	95% Confidence Interval	
			Lower	Upper
Index "perceived degree of innovation"	0.460	0.837	0.523	1.341
Index "anticipated impact"	0.000	11.486	5.797	22.759

Dependent Variable: Grade overall assessment
Model: (Threshold), index "perceived degree of innovation", index "anticipated impact", applicant's gender, applicant's age, applicant's academic degree, discipline group, type of institution, country of residence of external reviewer, source of recommendation, expertise of external reviewer.

Model fit	
Nagelkerke's Pseudo R^2	0.778
Likelihood-ratio test	0.000

This table only shows the values for the variables of interest. See Annex 4 for a full table.

Model 1.1a showed that almost all the components of the indices had a statistically significant effect. These were the aspects "research goal / topic", "approach / method" and the impact types "impact on patient / treatment", "impact / advances in the academic field" and "other impact". Only the mean value of the comments on the project in general had no statistically significant effect. From all the different components of the two indices, the types of impact were associated with larger increases in the odds of getting a higher degree. The "other impact" was associated with the largest increase in the odds of getting a higher grade, followed by "impact on patient / treatment" and "impact / advances in the field". The aspect "research goal / topic" was associated with a much lower increase in the odds of getting a higher grade, whereas the aspect "approach / method" was associated with a decrease in the odds of getting a higher grade.

However, this result should be treated with caution, especially because in the case of the different types of impact, the 95% confidence intervals were large. Large confidence intervals indicate that the odds ratios have a rather low level of precision (cf. Table 10, Szumilas, 2010, p. 227).

Table 10. Model 1.1a: extract of results

	Sig.	Odds Ratio	95% Confidence Interval	
			Lower	Upper
Aspect "project in general"	0.498	1.302	0.607	2.796
Aspect "research goal / topic"	0.003	5.515	1.812	16.789
Aspect "approach / method"	0.004	0.241	0.091	0.640
Impact type "impact on patient / treatment"	0.000	19.538	3.934	97.034
Impact type "impact / advances in the academic field"	0.000	7.796	2.963	20.508
Impact type "other impact"	0.000	265.816	29.522	2393.430

Dependent Variable: Grade overall assessment
Model: (Threshold), aspect “project in general”, aspect “research goal / topic”, aspect “approach / method”, impact type “impact on patient / treatment”, impact type “impact / advances in the academic field”, impact type “other impact”, applicant’s gender, applicant’s age, applicant’s academic degree, discipline group, type of institution, country of residence of external reviewer, source of recommendation, expertise of external reviewer.

Model fit	
Nagelkerke's Pseudo R^2	0.816
Likelihood-ratio test	0.000

This table only shows the values for the variables of interest. See Annex 4 for a full table.

The overall omnibus statistical tests showed the following variables to have a statistically significant effect in both models: the applicant’s gender, the discipline group and the source of recommendation ($p < 0.01$)[20]. For female applicants the odds of getting a higher grade, holding all else constant, are lower than for male applicants (OR = 0.011, 95% CI 0.002 to 0.051, $p < 0.0005$). The odds of getting a higher grade are not the same among the different discipline groups, $p < 0.0005$. Also the fact of being evaluated by a reviewer proposed by the SNSF, holding all else constant, decreased the odds of getting a higher grade compared to when the applicant himself proposed the reviewer (OR = 0.002, 95% CI 0.000 to 0.015, $p < 0.0005$). The applicant’s age, the academic degree, the type of institution and the country of residence of the external reviewer had no statistically significant effect on the prediction of the grade in both models ($p > 0.05$). Results were less robust for the expertise of the external reviewer. It had statistically significant effect in Model 1a ($p = 0.012$) but not in Model 1.1a ($p = 0.882$, cf. Annex 4 Tables 6.9-6.12).

[20] The numbers from Model 1.1a are reported, as confidence intervals are smaller than in Model 1a, which indicates a higher level of precision. For the full table of the parameter estimates please refer to Annex 4.

Hypothesis 1b

Again, the goodness-of fit tests could not be used because of a high number of cells with zero frequencies. The Nagelkerke coefficient amounted to 0.610, which means that Model 1b explained 61% of the variance. The likelihood-ratio test was significant too. This means that the variables entered add statistically significantly to the model estimated (cf. Table 11).
Model 1b was weaker than Model 1a and Model 1.1a when it came to predicting the category of the ordinal dependent variable correctly. It only predicted 52% of cases correctly. It was especially weak in predicting higher grades, most probably because there were fewer than in the case of the external reviews (cf. Annex 4 Table 6.13).

The results of the overall omnibus statistical tests showed that between the two variables of interest in Model 1b both had a statistically significant effect on the prediction of the grade given by the referees. An increase in either of them was associated with an approximately equal increase in the odds of getting a higher grade (cf. Table 11). Based on this result, hypothesis 1b is rejected.

Table 11. Model 1b: extract of results

	Sig.	Odds Ratio	95% Confidence Interval	
			Lower	Upper
Index "perceived degree of innovation"	0.000	2.208	1.440	3.386
Index "anticipated impact"	0.001	2.581	1.465	4.544

Dependent Variable: Grade comparative ranking
Model: (Threshold), index “perceived degree of innovation”, index “anticipated impact”, applicant’s gender, applicant’s age, applicant’s academic degree, discipline group, type of institution.

Model fit	
Nagelkerke's Pseudo R^2	0.610
Likelihood-ratio test	0.000

This table only shows the values for the variables of interest. See Annex 4 for a full table.

The overall omnibus statistical tests showed the following variables to have a statistically significant effect: the applicant’s age, the applicant’s academic degree and the discipline group ($p < 0.01$). The odds of getting a higher grade in a recommendation are not the same among the different age cohorts ($p = 0.003$) and the discipline groups ($p < 0.0005$). However, other than with external reviews there is no statistically significant difference between female and male applicants ($p = 0.410$), but there is a difference between applicants with a PhD and applicants who hold the title of a professor. Only having a PhD, holding all else constant, decreased the odds of getting a higher grade compared to applicants who hold a professorship (OR = 0.301, 95% CI 0.121 to 0.749, $p = 0.010$). Again the type of institution had no statistically significant effect on the prediction of the grade given by the referee (p = 0.662, cf. Annex 4 Tables 6.14-6.15).

6.2 Influence of the expressed perception of innovation on the probability of being funded

The relationship postulated in hypothesis 2 was tested using a binomial logistic regression analysis. With a binomial logistic regression the probability that an observation falls into one of the two categories of a dichotomous dependent variable can be predicted based on one or more independent variables that are either continuous or categorical. The goal is to investigate what effect the independent variables have on the probability that an event occurs (Bühl, 2010, p. 418; Lærd Statistics, 2015a).

6.2.1 Description of the model used to test H2

Figure 5 illustrates the model that was used to test hypothesis 2. It postulates the following relationship: among applications that were classified by the referees into the quality level where the funding line is drawn, the more an application is perceived as innovative, the smaller its chance of getting funding.

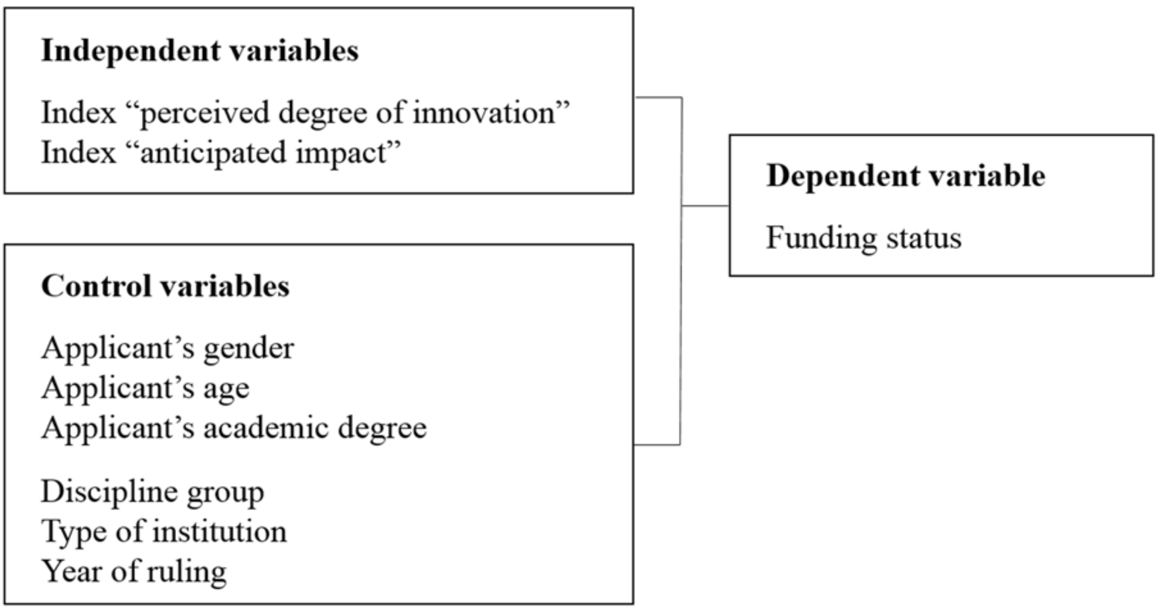

Figure 5. Illustration of Model 2

Statistical hypothesis

H_0: $OR_{innov} \geq 1$ and $OR_{impact} \geq 1$ **H_1:** $OR_{innov} < 1$ and $OR_{impact} < 1$

It was anticipated that the indices "perceived degree of innovation" and "anticipated impact" which were used as independent variables would have a negative effect on the likelihood that the application was approved.

The applicant's characteristics and the information related to the application[21] were entered as control variables to prevent overestimating or underestimating the effect of the independent variables.

6.2.2 Requirements of a binomial logistic regression

In order to perform a binomial logistic regression analysis and to obtain unbiased and efficient estimates some premises need to be fulfilled. The most important assumptions are listed below (Grüner, 2010; Lærd Statistics, 2015a). It is explained how the existing data met them and what further steps were taken.

1) *Dichotomous dependent variable.*

   The dependent variable was the funding status of the application – approved or rejected.

2) *Independent variables measured on a continuous or nominal scale.*

   The independent variables included in the model met this requirement.

3) *Independent observations and mutually exclusive and exhaustive categories of the variables.*

   There was no relationship between the categories of the variables and also no relationship between the categories either. Furthermore a case could not fall into more than one category of a variable, e.g. the application could not be attributed to two discipline groups at the same time.

4) *At least 10-15 cases per independent variable or a minimum of 100 cases.*

   In total there were 60 cases in the final model, which means that reliability of estimates might not be given. The model was still estimated in order to test whether the variables of interest have a significant influence.

   The variables "discipline group" and "type of institution" had a very low number of counts (<5), which is why these variables were not included in the logistic regression analyses. The

[21] Unlike in Model 1a and Model 1b the year of ruling was also entered as a control variable. This is because it is anticipated that it has an influence on the decision, but that the grades remain uninfluenced by this.

variable “applicant’s age” was transformed such that the categories 51-55 yrs, 56-60 yrs and 60+ yrs, with very low counts, were merged into one category (50+ yrs). This means that finally only four of the six control variables were entered into the calculation. No model with the components of the indices was estimated, as this would have meant including more independent variables thus violating this assumption even further.

5) *Linear relationship between the continuous independent variables and the logit transformed dependant variable.*
   The continuous independent variables need to be transformed into their natural logs to test this assumption. The two continuous independent variables (“perceived degree of innovation” and “anticipated impact”) contain a lot of zero values, which would produce missing values, as the natural log of zero is not defined. Therefore the continuous independent variables were transformed into easily interpretable dichotomous variables. As a result, this assumption was no longer relevant, as only nominal independent variables were used.
6) *No multicollinearity.*
   The VIF-values were smaller than 10 for the model estimated (cf. Annex 4 Table 6.16).
7) *No significant outliers, leverage or influential points.*
   There were five cases with studentized residual values greater than 2.5. They were inspected in detail in order to determine why they were outliers. Four of the five cases were outliers because they were rejected even though the index “perceived degree of innovation” was positive. They were not removed from the analysis.

6.2.3 Results from the binomial logistic regression analysis

The binomial logistic regression analysis was calculated using the logistic regression procedure in IBM SPSS Statistics 22. Again different methods to assess the model fit and their results will be explained. Thereafter, the parameter estimates for the independent variables will be presented and interpreted.

Overall, the model is statistically significant. The Hosmer and Lemeshow test is not statistically significant, which also indicates that the model is not a poor fit. When neither independent nor control variables are included, 73% of cases could be correctly classified. This value rises to 84% when all the variables are added. The addition of the independent and control variables improves the prediction of the cases into the observed categories of the dependent variable (cf. Table 12).

The only variable having a statistically significant influence on the funding status of an application is the index “perceived degree of innovation”. Unlike postulated in the hypothesis having a positive value in the index “perceived degree of innovation” it was associated with an increased likelihood of being funded compared to applications for which this index was zero or negative. However, due to the low number of cases included reliability of estimates might not be given. Based on this result, hypothesis 2 is rejected as well.

Table 12. Model 2: results

	Sig.	Odds Ratio	95% Confidence Interval	
			Lower	Upper
Index "perceived degree of innovation": positive (compared to zero or negative values)	0.030	8.217	1.233	54.774
Index "anticipated impact": positive (compared to zero or negative values)	0.331	0.301	0.027	3.387
Applicant's gender: females (compared to males)	0.784	0.730	0.077	6.939
Applicant's age (global variable)	0.130			
Applicant's age: 36-40 yrs (compared to -35 yrs old)	0.962	0.911	0.019	43.230
Applicant's age: 41-45 yrs (compared to -35 yrs old)	0.508	0.277	0.006	12.362
Applicant's age: 46-50 yrs (compared to -35 yrs old)	0.219	0.073	0.001	4.742
Applicant's age: 50+ yrs (compared to -35 yrs old)	0.040	0.012	0.000	0.810
Applicant's academic degree: PhD compared to professor	0.582	1.774	0.230	13.671
Year of ruling (global variable)	0.194			
Year of ruling: 2013 (compared to 2012)	0.192	0.086	0.002	3.429
Year of ruling: 2014 (compared to 2012)	0.396	0.259	0.011	5.854
Year of ruling: 2015 (compared to 2012)	0.209	0.139	0.006	3.008
Year of ruling: 2016 (compared to 2012)	0.025	0.010	0.000	0.561
Constant	0.112	62.700		

Dependent Variable: Funding status

Model fit	
Significance	0.020
Hosmer and Lemeshow Test	0.264
Nagelkerke's Pseudo R^2	0.480
Percentage accuracy in classification (PAC)	0.836
PAC without any independent variables	0.734

7 Conclusion

In this chapter the approach will be summarized. Then, the most important results will be analysed in light of the literature presented at the beginning. After that, the research question will be answered. The chapter will be concluded by a presentation of the implications for the SNSF and possible avenues for further research.

7.1 Summary of the approach

The present thesis attempted to assess the effectiveness of the evaluation procedure of the SNSF in terms of its ability to support innovative projects. To do this, the content of a randomly selected sample of external reviews and referees' recommendations that were provided in the time period from 2012 to 2016 to division biology and medicine were analysed. The evaluations were searched for statements including terms that were identified as pointing to innovation. By doing so the applications' degree of innovation as expressed by the reviewers could be determined.

After thorough descriptive analyses the hypotheses were tested using two different inferential statistical methods. In order to test hypothesis 1a / 1b ordinal logistic regression analyses were used. Hypothesis 2 was tested using a binomial logistic regression analysis.

7.2 Results

Building on the findings presented so far, the following can be summarized with respect to the evaluation procedure in division biology and medicine:

Main findings

- Based on existing literature, this thesis tried to prove that there is a negative relationship between the perceived degree of innovation of an application and the grade it is given by either the external reviewer or the referee. However, no such effect could be verified. On the contrary, in the case of external reviews there seems to be an inverse relationship between the anticipated impact and the grade. The more an application is perceived to have a high future impact by the external reviewer, the higher the probability that it receives a high grade.

  In a second step, it was attempted to prove that there is also a negative relationship between the perceived degree of innovation of an application and the grade given by the referee. Yet again no such effect could be verified. In contrast, in the case of the referees' recommendations there seems to be an inverse relationship between the perceived degree of innovation and the anticipated impact and the grade. The more an application is perceived to be innovative and/or to have a high future impact by the referee, the higher the probability that it receives a high grade.

This means that both external reviewers and referees are able to adequately translate their perception of an application's degree of innovation into a grade. It seems that neither private interests nor their intellectual boundaries prevent them from reducing the information asymmetry for the funding agency.

- In the last step I tried to prove that applications that are perceived as more innovative have a smaller chance of getting funded than applications perceived as less innovative. Yet again, this anticipated effect could not be verified. Rather, if the perceived degree of innovation was positive, chances of being funded increased compared to a negative or zero value.

  It seems that the circumstances in the decision-making process of the NRC's division of biology and medicine do not prevent the consideration of research policy objectives. Even though decisions are reached based on majority votes, applications perceived innovative had higher chances of success.

Further insights

- Comments about the project in general do not have an influence on the grade given by the external reviewer. Only more specific comments impact the grade. Most interestingly a positive evaluation of the innovativeness of the approach / method reduces the chances of getting a higher grade by the external reviewer. It seems that innovation in this area is not rewarded. This might be due to the fact that it is hard for reviewers to adequately judge an approach they do not know themselves[22].
- It has been shown that there might be a gender bias in external peer review. Other things held constant, female applicants have a lower probability of getting a higher grade by the external reviewer than male applicants. This relationship could not be found in case of the referees' recommendations.

[22] Due to data limitations it was unfortunately not possible to perform this analysis for the referees and their recommendations as well. It would have been very interesting to see whether the results would have differed or been similar.

- External reviewers originating from an applicant's list tend to give higher grades than reviewers who were chosen by the SNSF. This shows that the SNSF made the right decision in abandoning this practice, eliminating a possible source of bias.
- The type of institution does not have an influence on the grades given by either the external reviewers or the referees.
- The results suggest that the academic degree plays a role in the review by the referees. Applicants holding a PhD, ceteris paribus, have a lower probability of getting a higher grade by the referees than do professors. However, this could also just be a result of the applicants' experience in grant writing, which is reflected in the quality of the application. It would be interesting to investigate whether there are any notable differences between applications from PhD holders and those from professors.

7.3 Answering the research question

The research question derived in the first chapter reads as follows: Do peer reviewers disadvantage innovative applications in the division biology and medicine? This has been shown not to be the case innovative applications were not disadvantaged. This indicates that the SNSF lives up to its aspiration to create an environment allowing for innovative ideas to be pursued through project funding (SNSF, 2015d, p. 4).

It must be noted that the findings presented here do not necessarily mean that the most innovative projects are encouraged through the SNSF's project funding[23]. With the method used it is uncertain whether the most innovative projects were really identified. It could be that only the small-scale innovations were captured, as it might be more difficult for reviewers to properly identify large-scale innovations. Furthermore, there is still the possibility that researchers submitting applications adapted their behaviour and avoid proposing innovative research projects at all as has been shown by Laudel (2006). This self-selection bias has to be borne in mind but there does not seem to be a possibility to properly account for it in a study like this one. Another problem that could potentially be resolved by using an objective measure as the ones presented in chapter 4, is that not mentioning some reasons in an evaluation does not imply that those reasons have not been considered (Neidhardt, 1988, p. 94). This means that it cannot be excluded that the reviewers considered the factor innovation in their grading without mentioning it in their written messages.

[23] The SNSF has a particular funding scheme – Sinergia - aiming to encourage breakthrough research, which should attract the large-scale innovations.

7.4 Implications for the SNSF

In general the results are reassuring for the SNSF. There does not seem to be a problem with the support for innovative projects. However, reviewers usually assess the innovation potential only very briefly. Therefore, if the SNSF wanted to obtain a more in-depth evaluation of an application's potential for innovation an adaption of the guidelines for external reviewers and referees specifying what is expected in this regard in more detail could be beneficial. Most of the external reviewers seem to have followed the guidelines quite closely, so their behaviour could be guided in the desired direction.

There are however, some indications that the judgements made in peer review might be biased. Specially, a gender bias cannot be ruled out. So far, the SNSF monitors the differences between the success rate of female and male applicants, but it might also be advisable to monitor differences in external peer review.

As in peer review in general, and at the SNSF in particular, there are other potential sources of bias and inefficiency, it seems to be warranted to look into modifications to the peer review process or even alternatives to peer review.

7.4.1 Modification to the peer review process

Two diametrically opposed schools of thought can be distinguished when it comes to modifying the peer review process: blinding vs. opening up.

Blinding peer review

One possible modification to the grant peer review process could be to blind the reviewers to applicants and their institution. However, this has proven difficult to achieve. There is a large amount of material to judge a grant proposal by. At the SNSF for example, the submitted research plan can be up to 20 pages long. Furthermore, research plans usually include self-citations, which facilitates an understanding of who is behind an application. Yet, the greater problem would be that the track record and expertise of the applicants is an official evaluation criterion at the SNSF and this cannot be judged without knowing who submitted the application (SNSF, 2017; Wessely, 1998, p. 303). For these reasons, it seems that blinding peer review at the SNSF would be difficult.

Open peer review

The opposite solution would be to open the process up and move into the direction of open peer review. So far there is no common definition of open peer review, but it normally starts with providing the applicant with signed reviews and can go so far as to turning the evaluation procedure into a public process to which everyone can contribute (Ford, 2013, pp. 313–315).

The question is whether open peer review can solve the problems which surface in traditional peer review. One of the advantages of open peer review is that it would motivate reviewers to write better reviews. Also the identification of conflicts of interests would be simplified and abuse could be reduced. In that open peer review also allows an exchange between the applicant and the reviewer, this would also help to improve applications and finally the research carried out. A possible disadvantage is that finding enough reviewers could become even more difficult. Especially for younger, less established researchers it may be detrimental to openly criticise applications from others. According to those opposed to open peer review, the anonymity in the traditional system protects reviewers from such sanctions. Another potential problem is the confidentiality of the proposed research. Making a research plan publicly available could possibly encourage others to implement the proposed project (de Magalhães, 2012; Ford, 2013, pp. 317–320; Nature Editorial, 1999, pp. 197–198; The GovLab, 2016).

In principle, open peer review is a modification of the current procedure and one that the SNSF should look into in the near future, as it may help to improve the quality and transparency of the review process and thus increase the legitimacy of the organisation as a whole. Today the applicants have access to the anonymised versions of the reviews regarding their application. Article 3 in the RIPA states that the identity of the reviewers can only be unveiled with their consent, in case of a complaint. This means that the law would most possibly have to be amended in order to allow for signed reviews to be implemented.

7.4.2 Alternatives to peer review

There are some alternatives to peer review in grant decision-making. The most well known alternative is to make funding decisions based on quantitative measures of research output. This is also referred to as bibliometrics or scientiometrics. They are post hoc measures, meaning that they only reflect past performance and give no direct indication of the potential of a proposed project. It is based on past peer review and peer attention, as it looks at the number of accepted journal articles and their respective citation counts. Therefore, by using bibliometrics funding agencies could make use of the effort that has already been made in the journal peer review process. Thus, reducing the current inefficiency in grant peer review. However, as long as journal peer review is not free of bias, funding allocation based on bibliometrics would just be replicating these biases. Further, one of the main criticisms against bibliometrics is that focusing on such quantitative measures creates disincentives, ultimately leading to non-reproducible science. Lastly, as the measures obtained determine quantity, they are at best proxies of scientific excellence. In conclusion it seems unrealistic that bibliometrics can replace grant peer review entirely. It seems much more likely that such quantitative measures can and will be used as complementary information when evaluating grant proposals (ESF, 2011, p. 25; Langfeldt & Kyvik, 2011, p. 207; Lauer & Nakamura, 2015, pp. 1893–1894;Wessely, 1998, p. 304).

Another possible alternative would be to award grants at random. So far, this has already been implemented for small grants by some funding agencies. An option would be to allocate all the funding available randomly via a lottery, which is inherently unbiased. Alternatively, random allocation could be reserved for proposals that are close to the funding line. However, to determine which proposals fall into which group, peer review would still be required, but it could possibly be limited to filtering out the really bad projects. The benefits of such an approach are that funding would be independent of the applicants' reputation, therefore reducing bias, but with the risk that possibly not the best proposals are funded (Avin, 2017; Guthrie, Guérin, Wu, Sharif, & Wooding, 2013).

Artificial intelligence represents the most futuristic alternative to peer review. The recent developments and successes achieved in the field of artificial intelligence make it seem more plausible that the humans involved in the peer review process could be replaced by artificial intelligence in the future. Some authors think that learning algorithms could manage the whole review process by drawing on the enormous masses of data available in funders' and publishers' databases. However, up until now, understanding and handling human language still represents a major challenge for artificial intelligence and it still seems a long way until artificial intelligence could be able to take over the grant review process. Therefore, possibly a more readily implementable way of using artificial intelligence would be to use it for automated checks for plagiarism and conflicts of interest, and for the handling of the communication between funders, applicants and reviewers (Hukkinen, 2017; Stockton, 2017).

The major drawback of these possible alternatives from the point of view of the funding agencies is that it would reduce their credibility and legitimacy. However, as has been shown in chapter 2.2.2 peer review is predominant in science mainly because it serves the interests of the scientific community and the policy makers, and legitimates funding decisions to both of them (König, 2015, p. 3). Any real alternative would have to be able to do the same in order to be accepted by all the actors involved.

7.5 Possibilities for further research

There are several possibilities for further research in this area. First of all it would be insightful to explore the functioning of the NRC's division of biology and medicine for example with the help of participatory observation and relate those findings to the results obtained in this study. Another way to gain more insight into these processes would be to interview the different actors involved in the peer review process at the SNSF. In a next step, research could also be extended to the other two divisions of the SNSF.

Another possible future avenue for research could be to compare the results obtained here with a more objective measure of the degree of innovation of an application, for example, with one of the approaches briefly explained in chapter 4.

8 Bibliography

Abdoul, H., Perrey, C., Amiel, P., Tubach, F., Gottot, S., Durand-Zaleski, I., & Alberti, C. (2012). Peer Review of Grant Applications: Criteria Used and Qualitative Study of Reviewer Practices. *PLoS ONE*, *7*(9). http://doi.org/10.1371/journal.pone.0046054

AHRQ - Agency for Healthcare Research and Quality. (2014). An overview of the peer review process for AHRQ grant application submissions. Retrieved October 3, 2016, from http://www.ahrq.gov/funding/process/review/peerproc.html

Arrow, K. J. (1985). The Eonomics of Agency. In J. W. Pratt & R. J. Zeckhauser (Eds.), *Principals and Agents: The Structure of Business* (pp. 37–51). Boston, Massachusetts: Harvard Business School Press.

Avin, S. (2017). Science funding is a gamble so let's give out money by lottery. Retrieved March 27, 2017, from https://aeon.co/ideas/science-funding-is-a-gamble-so-lets-give-out-money-by-lottery

Baccarini, D., & Melville, T. (2011). Risk management of research projects in a University context - An exploratory study. Conference Paper, Goald Coast, Australia: 36th Australasian University Building Educators Association (AUBEA) Conference.

Berezin, A. (1998). The perils of centralized research funding systems. *Knowledge, Technology & Policy*, *11*(3), 5–26. http://doi.org/10.1007/s12130-998-1001-1

Böhmer, S., Neufeld, J., Hinze, S., Klode, C., & Hornbostel, S. (2011). *Wissenschaftlerbefragung 2010: Forschungsbedingungen von Professorinnen und Professoren an deutschen Universitäten* (iFQ-Working Paper No. 8). *iFQ-Working Paper*. Bonn. Retrieved from http://www.forschungsinfo.de/Publikationen/Download/working_paper_8_2010.pdf

Borlaug, S. B. (2015). Moral hazard and adverse selection in research funding : Centres of excellence in Norway and Sweden. *Science and Public Policy*, 1–11. http://doi.org/10.1093/scipol/scv048

Bornmann, L. (2011). Scientific Peer Review. *Annual Review of Information Science and Technology*, *45*(1), 197–245. http://doi.org/10.1353/pla.2004.0033

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance and Resource Allocation in Science. *Management Science*, (May), forthcoming. http://doi.org/10.1287/mnsc.2015.2285

Braun, D. (1993). Who Governs Intermediary Agencies? Principal-Agent Relations in Research Policy-Making. *Journal of Public Policy*, *13*(2), 135–162. http://doi.org/10.1017/S0143814X00000994

Braun, D. (2003). Lasting tensions in research policy-making - a delegation problem. *Science and Public Policy*, *30*(5), 309–321.

Braun, D. (2006). Delegation in the distributive policy arena: The case of research policy. In F. Gilardi & D. Braun (Eds.), *Delegation in Contemporary Democracies* (pp. 146–170). London: Routledge. http://doi.org/10.4324/9780203306512

Braun, D., & Guston, D. H. (2003). Principal-agent theory and research policy: an introduction. *Science and Public Policy*, *30*(5), 302–308. http://doi.org/10.3152/ 147154303781780290

Bühl, A. (2010). *PASW 18 Einführung in die moderne Datenanalyse* (12th ed.). München: Pearson Studium.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 21–30). New York: Cambridge University Press. http://doi.org/10.1017/CBO9780511816796.002

Chubin, D. E. (1994). Grants Peer Review in Theory and Practice. *Evaluation Review*, *18*(1), 20–30. http://doi.org/10.1177/0193841X9401800103

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, Massachusetts: The Belkap Press of Harvard University Press.

Coryn, C. L. S., Applegate, E. B., Schröter, D. C., Martens, K. S., & McCowen, R. H. (2012). *An Evaluation of the Transparency and Overall Quality of Evaluation at the Swiss National Science Foundation: Final Report*. Kalamazoo, MI.

de Magalhães, J. P. (2012). Open Peer Review. Retrieved August 28, 2016, from http://jp.senescence.info/thoughts/open_peer_review.html

Demicheli, V., & Pietrantonj, C. (2007). Peer review for improving the quality of grant applications. *Cochrane Database of Systematic Reviews*, (2), Art. No. MR000003. http://doi.org/10.1027/2151-2604/a000103

Diekmann, A. (2005). *Empirische Sozialforschung - Grundlagen, Methoden, Anwendungen* (14. Auflag). Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Dirk, L. (1999). A Measure of Originality: The Elements of Science. *Social Studies of Science*, *29*(5), 765–776. http://doi.org/10.1177/030631299029005004

Eisenhardt, K. M. (1989). Agency Theory: An Assessment and Review. *Academy of Management Review*, *14*(1), 57–74.

ESF - European Science Foundation. (2011). *European Peer Review Guide: Integrating Policies and Practices into Coherent Procedures*. Strasbourg.

Ford, E. (2013). Defining and Characterizing Open Peer Review: A Review of the Literature. *Journal of Scholarly Publishing*, *44*(4), 311–326. http://doi.org/10.3138/jsp.44-4-001

Früh, W. (2015). *Inhaltsanalyse* (8. Auflage). Konstanz und München: UVK Verlagsgesellschaft mbH.

Gallo, S. A., Sullivan, J. H., & Glisson, S. R. (2016). The Influence of Peer Reviewer Expertise on the Evaluation of Research Funding Applications. *PLoS ONE*, *11*(10), 1–18. http://doi.org/10.1371/journal.pone.0165147

Gigerenzer, G., & Selten, R. (2002). Rethinking rationality. In G. Gigerenzer & R. Selten (Eds.), *Bounded Rationality: The Adaptive Toolbox* (First, pp. 1–12). Cambridge, Massachusetts: Massachusetts Institute of Technology.

Gilardi, F., & Braun, D. (2002). Delegation aus der Sicht der Prinzipal-Agent-Theorie. *Politische Vierteljahresschrift*, *43*(1), 147–161.

Grüner, H. (2010). Tutorial zur Regressionsanalyse VII - Logistische Regression. Retrieved March 7, 2017, from http://gruener.userpage.fu-berlin.de/tutorials/logistische_regression.htm

Gulbrandsen, M. (2005). Tensions in the research council-research community relationship. *Science and Public Policy*, *32*(3), 199–209. http://doi.org/10.1038/315710b0

Guston, D. H. (1996). Principal-agent theory and the structure of science policy. *Science and Public Policy*, *23*(4), 229–240.

Guston, D. H. (2000). *Between Politics and Science: Assuring the Productivity and Integrity of Research*. Cambridge, United Kingdom: Cambridge University Press.

Guthrie, S., Guérin, B., Wu, H., Sharif, I., & Wooding, S. (2013). *Alternative to Peer Review in Research Project Funding* (Vol. RR-139-DH). Retrieved from http://www.rand.org/pubs/technical_reports/2009/RAND_TR608.pdf

Gutknecht-Gmeiner, M. (2008). *Externe Evaluierung durch Peer Review: Qualitätssicherung und -entwicklung in der beruflichen Erstausbildung* (1. Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hackett, E. J., & Chubin, D. E. (2003). *Peer Review for the 21 st Century : Applications to Education Research*. *Prepared for a National Research Council Workshop*. Washington, D.C.

Hartmann, I., & Neidhardt, F. (1990). Peer review at the Deutsche Forschungsgemeinschaft. *Scientometrics*, *19*(5–6), 419–425. http://doi.org/10.1007/BF02020704

Hollanders, H., Es-Sadki, N., & Kanerva, M. (2015). *Innovation Union Scoreboard 2015*. Retrieved from https://www.kowi.de/Portaldata/2/Resources/fp/2015-Innovation-Union-Scoreboard-Report.pdf

Horrobin, D. F. (1990). The Philosophical Basis of Peer Review and the Suppression of Innovation. *JAMA: The Journal of the American Medical Association*, *263*(10), 1438–1441. http://doi.org/10.1001/jama.263.10.1438

Horrobin, D. F. (1996). Peer review of grant applications: a harbinger for mediocrity in clinical research? *The Lancet*, *348*(9037), 1293–1295.

Hukkinen, J. I. (2017). Peer Review Has Its Shortcomings, But AI Is a Risky Fix. Retrieved March 31, 2017, from https://www.wired.com/2017/01/peer-review-shortcomings-ai-risky-fix/

Kaplan, D. (2005). How to Improve Peer Review at NIH. *The Scientist*, *19*(17), 10.

König, T. (2015). *Towards an analytical understanding of peer review in research funding*. Retrieved from https://www.researchgate.net/profile/Thomas_Koenig13/publication/279195198_Towards_an_analytical_understanding_of_peer_review_in_research_funding/links/558e3b8a08aed6ec4bf36de8/Towards-an-analytical-understanding-of-peer-review-in-research-funding.pdf

Lærd Statistics. (2015a). Binomial logistic regression using SPSS Statistics. Retrieved March 7, 2017, from https://statistics.laerd.com/premium/spss/blr/binomial-logistic-regression-in-spss.php

Lærd Statistics. (2015b). Ordinal logistic regression using SPSS Statistics. Retrieved March 3, 2017, from https://statistics.laerd.com/premium/spss/olr/ordinal-logistic-regression-in-spss.php

Lamont, B. M., Fournier, M., Guetzkow, J., Mallard, G., & Bernier, R. (2007). Evaluating creative minds : The assessment of originality in peer review. In A. Sales & M. Fournier (Eds.), *Knowledge, communication and creativity* (pp. 166–182). London: SAGE Publications. http://doi.org/10.4135/9781446215548.n10

Langfeldt, L. (2001). The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effect on the Review Outcome. *Social Studies of Science*, *31*(6), 820–841.

Langfeldt, L. (2006). The Policy Challenges of Peer Review: Managing Bias, Conflict of Interests and Interdisciplinary Assessments. *Research Evaluation*, *15*(1), 31–41. http://doi.org/10.3152/147154406781776039

Langfeldt, L., & Kyvik, S. (2011). Researchers as evaluators: Tasks, tensions and politics. *Higher Education*, *62*(2), 199–212. http://doi.org/10.1007/s10734-010-9382-y

Langfeldt, L., Ramberg, I., & Gunnes, H. (2014). *Researcher Survey for the Swiss National Science Foundation (SNSF)*. Nordic Institute for Studies in Innovation, Research and Education (NIFU), Oslo.

Laudel, G. (2006). The art of getting funded: how scientists adapt to their funding conditions. *Science and Public Policy*, *33*(7), 489–504. http://doi.org/10.3152/147154306781778777

Lauer, M. S., & Nakamura, R. (2015). Reviewing Peer Review at the NIH. *The New England Journal of Medicine*, *373*(20), 1893–1895. http://doi.org/10.1056/NEJMp1002530

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in Peer Review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. http://doi.org/10.1002/asi

Lee, Y., Chung, D., & Kim, Y. (2007). Public R&D Risk and Risk Management Policy. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, *1*(10), 508–511.

Leung, F., & Isaacs, F. (2008). Risk management in public sector research : approach and lessons learned at a national research organization. *R&D Management*, *38*(5), 510–519.

Lewandowsky, S., & Thomas, J. (2009). Expertise: Acquisition, limitations, and control. *Reviews of Human Factors and Ergonomics*, *5*(1), 140–165. http://doi.org/10.1518/155723409X448044.

Luppino, R., Hosseini, M. R., & Rameezdeen, R. (2014). Risk Management in Research and Development (R&D) Projects: The Case of South Australia. *Asian Academy of Management Journal*, *19*(2), 67–85.

Luukkonen, T. (2012). Conservatism and risk-taking in peer review: Emerging ERC practices. *Research Evaluation*, *21*(1), 48–60. http://doi.org/10.1093/reseval/rvs001

Maletta, H. (2007). Weighting. Unpublished work. Retrieved from www.spsstools.net/static /resources/WEIGHTING.pdf

Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H.-D., & O 'mara, A. (2009). Gender Effects in the Peer Reviews of Grant Proposals: A Comprehensive Meta-Analysis Comparing Traditional and Multilevel Approaches. *Review of Educational Research*, *79*(3), 1290–1326. http://doi.org/10.3102/0034654309334143

McCullough, J. (1989). First comprehensive survey of NSF applicants focuses on their concerns about proposal review. *Science, Technology & Human Values*, *14*(1), 78–88. http://doi.org/10.1177/016224398901400107

Merkhofer, M. W. (1987). *Decision Science and Social Risk Management* (2nd ed.). Springer Netherlands.

Merton, R. K. (1968). The Matthew Effect in Science. *Science*, *159*(3810), 56–63. Retrieved from http://www.jstor.org/stable/10.7312/rign14948

Mervis, J. (2014). Peering into peer review. *Science*, *343*(6171), 596–598. http://doi.org/10.1126/science.343.6171.596

Mitroff, I. I., & Chubin, D. E. (1979). Peer Review at the NSF: A Dialectical Policy Analysis. *Social Studies of Science*, *9*(2), 199–232. http://doi.org/10.1177/030631277900900203

MRC - Medical Research Council. (2016). Guidance for peer reviewers - A detailed guide for reviewers of MRC proposals, including assessment criteria and scoring systems. Guidelines. Retrieved from https://www.mrc.ac.uk/documents/pdf/reviewers-handbook/

Nature Editorial. (1999). Pros and cons of open peer review. *Nature Neuroscience*, *2*(3), 197–198. http://doi.org/10.1038/6295

Neidhardt, F. (1988). *Selbststeuerung in der Forschungsförderung: Das Gutachterwesen der DFG*. Opladen: Westdeutscher Verlag GmbH.

Neuendorf, K. A. (2002). *The Content Analysis Guidebook*. Thousand Oaks, California: SAGE Publications.

Neufeld, J., Hinze, S., & Hornbostel, S. (2014). *Befragung des wissenschaftlichen Personals an österreichischen Hochschulen und außeruniversitären Forschungseinrichtungen*. Berlin. Retrieved from https://zenodo.org/record/17854/files/iFQ-FWF-Umfrage-Bericht.pdf

NHMRC - National Health and Medical Research Council. (2015). Project Grants - Scheme-specific peer review guidelines. Attachment H: Guidance for assessors to assess the project grants assessment criteria. Retrieved from https://www.nhmrc.gov.au/_files_nhmrc/file/grants/apply/peer_review_guidelines/2015/project_grants_peer_review_guidelines_2015_attachment_h_150310.pdf

NIH - National Institutes of Health. (2016). Definitions of Criteria and Considerations for Research Project Grant (RPG/X01/R01/R03/R21/R33/R34) Critiques. Retrieved October 3, 2016, from https://grants.nih.gov/grants/peer/critiques/rpg_D.htm

NSERC - Natural Sciences and Engineering Research Council of Canada. (2016). Discovery Grants Peer Review Manual 2016-17. Manual.

O'Sullivan, D., & Dooley, L. (2009). *Applying innovation*. Thousand Oaks, California: SAGE Publications. Retrieved from http://www.oecd.org/site/innovationstrategy/defining innovation.htm\nhttp://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19721435

Reinhart, M. (2012). *Peer Review in der Forschungsförderung : das wissenschaftliche Begutachtungsverfahren des Schweizerischen Nationalfonds*. (A. Bora, S. Maasen, C. Reinhardt, & P. Wehling, Eds.) *Schriftenreihe "Wissenschafts- und Technikforschung" Band 10 Soziologie und Epistemologie des Peer Review* (1. Auflage). Baden Baden: Nomos Verlagsgesellschaft. http://doi.org/10.1017/CBO9781107415324.004

Rip, A. (1994). The republic of science in the 1990s. *Higher Education*, *28*(1), 3–23.

Rip, A. (2000). Higher Forms of nonsense. *European Review*, *8*(4), 467–486.

RIPA. Federal Act on the Promotion of Research and Innovation (RIPA) (2015). Berne: Bundesversammlung der Schweizerischen Eidgenossenschaft. Retrieved from https://www.admin.ch/opc/de/classified-compilation/20091419/index.html

Rössler, P. (2010). *Inhaltsanalyse* (2. Edition). Konstanz: UVK Verlagsgesellschaft mbH.

SBFI. (2016). Forschung und Innovation in der Schweiz. Retrieved May 26, 2016, from https://www.sbfi.admin.ch/sbfi/de/home/themen/forschung-und-innovation-in-der-schweiz.html

Scherngell, T., Roche, I., Hörlesberger, M., Besagni, D., Züger, M. E., & Holste, D. (2013). Initial comparative analysis of model and peer review process for ERC starting grant proposals. *Research Evaluation*, *22*(4), 248–257. http://doi.org/10.1093/reseval/rvt015

Schumpeter, J. A. (1939). *Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process*. New York Toronto London: McGraw-Hill Book Company. Retrieved from http://classiques.uqac.ca/classiques/Schumpeter_joseph/business_cycles/schumpeter_business_cycles.pdf

Sharif, I., Farrands, A., & Wooding, S. (2009). *Evaluating Grant Peer Review in the Health Sciences: A review of the literature*. Retrieved from http://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR742.pdf

Simonton, D. K. (2012). Taking the U.S. Patent Office Criteria Seriously: A Quantitative Three-Criterion Creativity Definition and Its Implications. *Creativity Research Journal*, *24*(2–3), 97–106. http://doi.org/10.1080/10400419.2012.676974

Slipersæter, S., Lepori, B., & Dinges, M. (2007). Between policy and science: Research councils' responsiveness in Austria, Norway and Switzerland. *Science and Public Policy*, *34*(6), 401–415. http://doi.org/10.3152/030234207X239456

SNSF. Statutes of the Swiss National Science Foundation (2002). Retrieved from http://www.snf.ch/SiteCollectionDocuments/statuten_08_e.pdf

SNSF. Organisational Regulations of the National Research Council of 14 November 2007 (2007). Retrieved from http://www.snf.ch/SiteCollectionDocuments/por_org_rec_reglement_e.pdf

SNSF. (2011a). Guidelines for the assessment of applications by members of the Research Council. Internal guideline.

SNSF. (2011b). Research for you - The National Research Programmes (NRPs). Berne. Retrieved from http://www.snf.ch/SiteCollection Documents/nrp_brochure_e.pdf

SNSF. (2013). Action Plan 2013-2016. Berne. Retrieved from http://www.snf.ch/SiteCollectionDocuments/snf_aktionsplan_2013_2016_e.pdf

SNSF. (2014). Guidelines for the assessment by external reviewers. Internal guideline.

SNSF. (2015a). 6th Monitoring Report on Peer Review. Internal Report.

SNSF. (2015b). Annual Report 2014. Berne. Retrieved from http://www.snf.ch/SiteCollectionDocuments/medienmitteilungen/jb_gesamtaus gabe_2014_e.pdf

SNSF. (2015c). Multi-Year Programme 2017 – 2020. Berne. Retrieved from http://www.snf.ch/SiteCollectionDocuments/mehrjahresprogramm_2017_2020_e.pdf

SNSF. (2015d). Profile - Facts and Figures. Berne. Retrieved from www.snf.ch/SiteCollectionDocuments/inb_por_d.pdf

SNSF. Regulations of the Swiss National Science Foundation on research grants (Funding Regulations) (2015). Ausschuss des Stiftungsrates. Retrieved from http://www.snf.ch/SiteCollectionDocuments/allg_reglement_16_e.pdf

SNSF. (2016a). 7th Monitoring Report on Peer Review. Internal Report.

SNSF. (2016b). Der SNF - Porträt - Strategie. Retrieved June 16, 2016, from http://www.snf.ch/de/derSnf/portraet/strategie/Seiten/default.aspx

SNSF. (2016c). Profile 2015-2016. Berne. Retrieved from http://www.snf.ch/SiteCollectionDocuments/jahresbericht_Profil2015-2016_gesamt_en.pdf

SNSF. (2016d). Statistics 2015 - Full version. Retrieved from http://www.snf.ch/SiteCollectionDocuments/statistiken_vollversion_en.pdf

SNSF. (2016e). The SNSF - Evaluation procedures. Retrieved April 18, 2016, from http://www.snf.ch/de/derSnf/auswahlverfahren/projektfoerderung/Seiten/default.aspx#

SNSF. (2017). Requirements for the research plan. Retrieved March 17, 2017, from http://www.snf.ch/en/funding/documents-downloads/Pages/regulations-requirements-for-the-research-plan.aspx

Spier, R. (2002). The history of the peer-review process. *TRENDS in Biotechnology*, *20*(8), 357–358.

SSIC. (2015). *Akteurskonstellationen im Schweizer Bildungs- , Forschungs- und Innovationssystem*. Retrieved from http://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR742.pdf

Sternberg, R. J., Pretz, J. E., & Kaufman, J. C. (2003). Types of Innovations. In L. V. Shavinina (Ed.), *The International Handbook on Innovation* (First edit, pp. 158–169). Oxford: Elsevier Science Ltd. http://doi.org/10.1016/B978-008044198-6/50011-5

Stockton, N. (2017). If AI Can Fix Peer Review in Science, AI Can Do Anything. Retrieved March 31, 2017, from https://www.wired.com/2017/02/ai-can-solve-peer-review-ai-can-solve-anything/

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie Canadienne de Psychiatrie de L'enfant et de L'adolescent*, *19*(3), 227–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20842279

The GovLab. (2016). Innovations in Granting: Open Peer Review and Participatory Judging. Retrieved July 28, 2016, from https://medium.com/open-grantmaking-innovations/innovations-in-granting-open-peer-review-and-participatory-judging-4bc48e904e41#.6bthlsogo

The Guidelines project. (2016). Innovation, originality. Retrieved July 14, 2016, from https://www.guidelines.kaowarsom.be/en/annex_4-6_innovation

Travis, G. D. L., & Collins, H. M. (1991). New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System. *Science, Technology & Human Values*, *16*(3), 322–341. http://doi.org/10.1177/016224399101600303

van der Meulen, B. (1998). Science policies as principal-agent games: Institutionalization and path dependency in the relation between government and science. *Research Policy*, *27*, 397–414. http://doi.org/16/S0048-7333(98)00049-3

van der Meulen, B. (2003). New roles and strategies of a research council: intermediation of the principal-agent relationship. *Science and Public Policy*, *30*(5), 323–336. http://doi.org/10.3152/147154303781780344

Wang, J., Lin, W., & Huang, Y. H. (2010). A performance-oriented risk management framework for innovative R&D projects. *Technovation*, *30*(11–12), 601–611. http://doi.org/10.1016/j.technovation.2010.07.003

Wang, J., Veugelers, M., & Stephan, P. (2016). *Bias Against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators* (NBER Working Paper Series No. 22180). Cambridge, Massachusetts. http://doi.org/10.3386/w22180

Wessely, S. (1998). Peer review of grant applications: what do we know? *The Lancet*, *352*(9124), 301–305.

Wood, F., & Wessely, S. (2003). Peer review of grant applications: a systematic review. In F. Godlee & T. Jefferson (Eds.), *Peer Review in Health Sciences* (Second edi, p. Kindle file). London: BMJ Books.

Annex 1. Coding Scheme

Introduction

The present study wants to analyse the SNSF's evaluation procedure. Hypotheses regarding its effectiveness will be tested. The analysis is restricted to selected topics.
The coding scheme is divided into several parts. First a part with formal codes, such as data related to the application and the responsible applicant, data related to the evaluation by the referee and the external reviewer. Second a part with the content elements, where really the degree of innovation as described by the evaluators will be assessed.

Formal identification codes

Data related to the application

Application Number (ApplNo)
CI: Enter the official number of the application.
Note: This number has to be changed to a random number before publication in order to guarantee anonymity.

Date of submission (DateSub)
CI: Enter the date of the application's submission as dd.mm.yyyy.

Date of decision (DateDecision)
CI: Enter the date the administrative offices sent out the decision letter as dd.mm.yyyy.

Year (Year)

1	2012
2	2013
3	2014
4	2015
5	2016
99	unable to determine / not knwon

Discipline group (DisGroup)

301 Basic Biological Research

302 General Biology

303 Basic Medical Sciences

304 Experimental Medicine

307 Clinical Medicine

308 Preventive Medicine (Epidemiology/Early Diagnosis/Prevention)

309 Social Medicine

Corresponds to the SNSF's official list of research domains and disciplines

Type of institution (InstType)

1 ETH domain (incl. ETH domain research institutes EMPA, EAWAG, PSI, WSL)

2 Cantonal universities

3 Universities of applied sciences and universities of teacher education

4 Others (research institutes, hospitals independent of higher education institutes,…)

99 unable to determine / not known

Use Inspired Application (UseInspiredDummy)

0 not use inspired

1 use inspired

Final rating as decided by NRC (FinalRatingNRC)

1 D

2 C

3 BC

4 B

5 AB

6 A

99 unable to determine / not known

Funding status (FundStatus)

0 negative funding decision (rejected)

1 positive funding decision (approved)

Data related to the responsible applicant

Gender of responsible applicant (Gender)

0 Male

1 Female

99 unable to determine / not known

Academic degree of responsible applicant (Degree)

1 No degree

2 Doctoral degree (incl. MD)

3 Professor

99 unable to determine / not known

Age of responsible applicant (Age)

1 up to 35 years old

2 36-40 years old

3 41-45 years old

4 46-50 years old

5 51-55 years old

6 56-60 years old

7 older than 60 years

99 unable to determine / not known

CI: to be calculated as age when the applicant submitted his application

→ up to here all data comes directly from the SNSF databases

Data related to the expert evaluating the application

Type of expert (ExpertType)

1 internal (referee) → *continue with section "recommendation by referee (internal expert)*

2 external → *continue with section "data on external review"*

Recommendation by referee (internal expert)

Scientific track record and expertise of the applicants (as a team)

(IR_GradeTrackRecord)

1 poor

2 average

3 good

4 very good

5 excellent

6 outstanding

99 unable to determine / not known

Assessment of the proposed project (with reference to the SNSF criteria) (IR_GradeProject)

1 poor

2 average

3 good

4 very good

5 excellent

6 outstanding

99 unable to determine / not known

Comparative ranking (IR_CompRating)

1 D: The proposal is among the weakest 10%

2 C: 75% of the proposals are stronger, 10% weaker

3 BC: 50% of the proposals are stronger, 25% weaker

4 B: 50% of the proposals are weaker, 25% stronger

5 AB: 75% of the proposals are weaker, 10% stronger

6 A: The proposal is among the strongest 10%

99 unable to determine / not known

→ *continue with section "perception of evaluation"*

Data on external review

Source of recommendation (ER_SourceRecommendation)

1 recommended by applicant

2 recommended by referee

3 recommended by the SNSF

4 recommended by a reviewer

5 recommended by referee OR the SNSF (only if this category has been selected)

99 unable to determine / not known

Expertise of the reviewer (ER_Expertise)

1 is within my area of specialisation

2 is within my wider discipline

Usefulness of the review (as judged by the internal expert) (ER_Usefulness)

1 useful

2 in part useful

3 not useful → *STOP CODING*

99 unable to determine / not known

External expert count (ER_Count)

Number from 1 to n

CI: Note how many external experts in total have written a review for a specific proposal.

Origin of external expert (ER_Origin)

1 Swiss

2 Other, non Swiss

99 unable to determine / not known

Evaluation by external reviewer

Applicants' scientific track record and expertise (ER_GradeTrackRecord)

1 poor
2 average
3 good
4 very good
5 excellent
6 outstanding
7 not considered
99 unable to determine / not known

Scientific relevance, originality and topicality (ER_GradeRelOrigTop)

1 poor
2 average
3 good
4 very good
5 excellent
6 outstanding
7 not considered
99 unable to determine / not known

Suitability of methods and feasibility (ER_GradeMethFeas)

1 poor
2 average
3 good
4 very good
5 excellent
6 outstanding
7 not considered
99 unable to determine / not known

Overall assessment (ER_GradeOverall)

1 poor
2 average

3 good
4 very good
5 excellent
6 outstanding
7 not considered
99 unable to determine / not known

Appearance of evaluation

Length of evaluation (Length)

CI: Enter the number of words that have been written.

CI: In case of the referee's recommendation the text of the following sections is to be considered:

- Applicants' scientific track record and expertise
- Assessment of the proposed project
- Comparative ranking

CI: In case of the external evaluation the text of the following sections is to be considered:

- Comments regarding the overall assessment
- Applicants' scientific track record and expertise
- Scientific relevance, originality and topicality (incl. broader impact)
- Suitability of methods and feasibility

Quality of the evaluation (Quality)

3 good
2 bad
99 unable to determine / not known

CI: Note the impression of the evaluation's quality.

CI: Indicators for bad evaluations are inappropriate, off-topic comments, repetition of the same parts in one section, accumulation of grammar mistakes, copy-paste of large parts from the external reviews without indication.

Perception of described degree of innovation (InnovationOverall)

6 extremely innovative

5 very innovative

4 quite innovative

3 somewhat innovative

2 not very / rather not innovative

1 not innovative

99 no mention of innovation

CI: Note the overall impression of the described degree of innovation.

Analysis of content (synthetic coding scheme)

General coding instructions:

CI: In case of doubt, the coding scheme is always to be consulted.

CI: Coding follows the flow of text in the evaluation. Statements that appear first are to be coded first.

CI: The text is to be searched for statements that fit into one of the categories of the variable <u>*term used to describe innovation.*</u> Only statements that describe the reviewer's perception/impression of project's innovativeness are going to be considered. Statements that do not articulate innovation in some way, reproduce what the applicant stated himself, repeats what has been stated in the application, the reviewer's own summary of the current state of the art and his suggestions for improvement are not considered. Example: *"doing xy would add to originality"* is not coded, because it is not a statement about the project's innovativeness directly but a suggestion for improvement. *"Doing xy adds to originality"* is coded, because here the xy is something the applicants intend to do. *"Proposal seeks to identify a new mechanism"* is not coded, because this is just a mere repetition of the aims as stated by the applicant.

CI: A statement is a grammatically complete expression, representing a semantic meaning on its own. Therefore a statement can equal a sentence, but it is possible that a sentence represents more than one statement. The sentence *"the project is original and innovative"* for example is built of two statements. First *"the project is original"* and second *"(the project is) innovative"*. The same category can be attributed more than once for the same sentence. As a rule of thumb it can be said that if there is a change in one of the three content variables a different statement is made. Of course also the second statement identified in this way still needs to be about innovation. This rule is going to be used in order to ensure reliability.

Examples:

- *"the present proposal is innovative using the newest technologies for obtaining the relevant tissue material by image guided biopsy taking."* In this case one would have to code two statements. The codes for the first statement are as follows 'innovative' + 'positive' + 'general' and for the second statement 'novel' + 'positive' + 'technology/technique'.
- *"Specific research aims are not very original, as the same kind of research has already been performed in the field of osteoarthritis".* In this case the following has to be coded: 'original' + 'negative' + 'research aims' and 'novel' + 'negative' + 'general'.
- Example where a sentence represents only one statement even though one of the three variables changes: *"Application takes a new approach by trying to find serum markers that are indicative for predisposition for infection"* In this case the information that is further provided can be used to specify the aspect described as being innovative. Code as 'novel' + 'positive' + 'research subject'.
- Example where a sentence represents only one statement about innovation even though one of the three variables changes: *"The objective of the study is innovative because of the paucity and non-randomized nature of the data currently available."* Code as 'innovative' + 'positive' + 'research aims'. Related example: *"Scientific originality is excellent, given that double-blind, well controlled studies are highly needed."* Code as 'original' + 'very positive' + 'approach'.
- Example where a sentence represents only one statement, because the same term is used twice: *"Idea to use ketogens to increase energy supply while reducing oxidative stress is not entirely new, has been already explored"* Here the term used to describe innovation remains the same, which is why it counts only as one statement: 'novel' + 'negative' +'research material'. Related example *"Some aspect of the project are original such as the microscope based platform for drug profiling and the engraftment of NSG mice with MRD samples":* the two aspects described are both related to methods, therefore it is going to be coded only once as 'original' + 'positive' + 'method'. Another example: *"the impact of this grant in the understanding of biological determinant of disease progression is of great importance clinically for a*

better management of patients treatment." The term used both times refers to improvements for patients / clinical relevance, therefore it is coded as follows: 'clinical relevance' + 'very positive' (because it is of great importance) + 'general'.

- When a statement in one part is repeated in another part, to make another argument, it is entered twice / as two statements. Example: *"Combination of MRS, PET and fMRi is original".* Suitability of methods: *"Despite the originality of combining MRS, PET and fMRI there is concern as …".*

CI: Only statements referring clearly to the proposed research project are to be considered. Examples of statement that should not be coded: *"forensic toxicology is per se user-oriented, results can be put into practice directly", "SPT, as a potential of new drugs in the treatment of metabolic syndrome, is of interest of today's biomedical research"* → the potential is described in general terms, with no reference to the proposed project. Examples of statements that has to be coded although (without considering the context) at first they seem not be related to the proposed research project: "*Thus, regenerative medicine of pancreatic beta-cells is an intriguing approach to development of future therapy for diabetes"* → 'clinical relevance' + 'positive' + 'approach'. *"Although many studies on islet cell plasticity have focused on inter-conversion between alpha-cell and beta-cell to date, there is no study on delta-cell conversion to beta-cell"*, which is what the proposed project set out to do → 'novel' + 'positive' + 'approach'.

CI: Before a statement is attributed to one category it has to be checked whether there is an alternative interpretation. If an alternative is present, the superordinate, more general category has to be used.

CI: If a statement fits into two categories of differing degrees of generality at the same time, the more specific category has to be chosen.

External evaluation	**Recommendation by referee**
CI: The part about the applicant's scientific track record and expertise is not to be analysed, neither text not belonging to the evaluation as such (further comments).	CI: The part about the applicant's scientific track record and expertise is not to be analysed, neither text not belonging to the evaluation as such (preliminary comments by the administrative offices, the outline of the proposed project and funding terms).

CI: The following content is to be coded: • comments regarding the overall assessment • scientific relevance, originality and topicality, plus broader impact if applicable (block of criteria) • topicality and the suitability of methods and feasibility (block of criteria).	CI: The following content is to be coded: • assessment of the proposed project • comments on the comparative ranking → if they refer to the external reviews or summarize what they have said in one of these parts it is also coded
CI: If the exactly same text appears completely in more than one of the text boxes belonging to the blocks of criteria (copy/paste) it is only assessed for the text box it appeared first. This text then does not enter into the page count.	Not applicable as only one block of criteria is analysed.
CI: If the comments regarding the overall assessment are composed of text snippets from the text boxes belonging to the blocks of criteria, the comments are going to be analysed. If entire text boxes are copy/pasted into the comments section, the comments are not to be analysed but only the text box. This is the exception to the rule that coding follows the flow of text. Additionally this text then does not enter into the page count either. This also applies to clearly distinguishable paragraphs in one text block (e.g. if the text block "scientific relevance, originality and topicality" has been split into three separate paragraphs).	CI: If the comments regarding the comparative ranking are composed of text snippets from the text boxes belonging to the blocks of criteria the comments are going to be analysed. If entire text boxes are copy/pasted into the comments section, the comments are not to be analysed but only the text box. This is the exception to the rule that coding follows the flow of text. Additionally this text then does not enter into the page count either. This also applies to clearly distinguishable paragraphs in one text block (e.g. if the text block "scientific relevance, originality and topicality" has been split into three separate paragraphs).

Number of statement (StatementNumber)

Enter the number of the statement starting from 1. If there is no statement about innovation enter 0.

Location of statement (Location)

This category captures the location of the statement, as the part of the evaluation where it is written down. In that sense it is a formal category. Statements from sources other than the categories listed are not to be coded (cf. general coding instructions).

10 Assessment of the proposed project (Referee's recommendation)
20 Comments regarding comparative ranking (Referee's recommendation)

60 Comments regarding the overall assessment (External evaluation)
70 Scientific relevance, orginality and topicality (External evaluation)
80 Suitability of methods and feasibility (External evaluation)

Term used to describe innovation (Term)

The terms listed in this category are mainly the terms that have been identified in the literature as being used synonymous or a description of the concept of innovation in research (cf. chapter 2.3).

10 innovative

This category does not require interpretation by the coder. Only the exact term (innovative, innovation) is coded. The classification to the category of valuation follows the standard rules.

20 original

This category requires some interpretation. Also synonymous descriptions such as ‘imagination has been used’ or ‘imaginative’ are attributed to this category. The classification to the category of valuation follows the standard rules.

30 novel / do sth that hasn't been done before

This category requires careful interpretation. If the statement includes the term ‘novel’ or ‘unprecedented’, classification to the category of valuation follows standard rules. The important thing here is to assure that the novel or unprecedented really refers to the proposed project. Statements describing how the proposed project or part of it is going to do something that has not been done before or the opposite that it is a simple replication of something that has already been done also belong to this category. When they write that there are already a lot of studies on that subject but it does not explicit that the current proposal just wants to replicate it is not coded. Examples of statement that should not be coded: *“The use of only novel oral* *anticoagulants in the study is well supported" or "Project relies on new tools"* → it doesn't say that it is a 'novel' idea to use these novel anticoagulants / new tools. It has to be clear that the project itself proposes new/novel methods only then can it be coded. *"Unresolved/neglected question / problem"* → this cannot be classified as novel or as something that has not been done before. It cannot be ruled out that some people have not tried to solve this question before the applicant. *"Double-blind, well controlled studies are highly needed".* → saying that there is a lack / paucity of studies of a certain type or that studies of a certain type are needed does not classify as novel / do sth that hasn't been done before. *"to date, data concerning the possible adverse effects of graphene are limited."* and *"Data on SIBO taking into account JHS status have not been published thus far."* → if it does not become evident from the context that the proposed project will lead to this data (which is not the case for the two examples), such statements are not coded. "*Scientific tasks proposed in this project are based on well known and published data."* → being based on well known / published / established / prior / exiting / preliminary data is not coded, as it is not clear whether this necessarily has to be seen as negative or not.

Specific rules for the classification to the category of valuation in case of statements describing something that has not been done before:

+1 statements' explaining how the project under evaluation is something that has not been done before are considered as positive Example: *"However, a similar strategy to unravel this problem as the one in the planned project has not been encountered previously." "So far nobody has examined ...", "They are the first to investigate", "It is the first time that ...".*

-1 Statements explaining that something has been done before are considered as negative. Examples: *"many of the questions have in some way already been addressed", "the idea to use xy is not entirely new and has already been explored*", *"proposed project appears to be redundant to published work", "this is a remake of a study from Canada".*

-2 Examples: *"Experiments that are simple repeats in mammalian cells of what was done on yeast are serious weaknesses of this project"* → the precision that this is a serious weakness accounts for the very negative valuation. *"There is nothing new in exploring xy", "lack of novelty", "Doing xy is not a new idea".*

Examples that should <u>not</u> be coded in this context: *"Only few research groups have taken efforts to shed some light on this huge problem"* is not related to innovation and at the same time it does not say that something has not been done before. "*The applicant want to use the same* *system to investigate the relevance of PU.1 and Gata1 during different lineage decisions or differentiation processes",* the applicant will not be doing something that has not been done before and at the same time from the context it is evident that the expert is neither being negative about it, which would mean that one should coded it as doing something that has been done before.

40 creative

This category requires some interpretation by the coder. Statements including the term unique are also assigned to this category, because creativity includes having unique ideas. Classification to the category of valuation follows the standard rules.

50 groundbreaking

The term requires some interpretation. Not only the term groundbreaking, but also closely related terms (e.g. breakthrough) are coded. The classification to the category of valuation follows the standard rules.

60 impact

This code is given if impact is described without any further specification or if the attribution to a subcategory is not unambiguous. Statements using the term *"broader impact"* with no further specification are generally coded using this category with a +1 valuation. If new/novel is used in a statement belonging to this category or one of its sub-categories it belongs to this category and not to category "30 novel". Example: *"May potentially offer development of novel therapies directly targeting disc-related back-pain."* The classification to the category of valuation follows in general the standard rules.

Specific rules for the classification to the category of valuation (also applicable to the subcategories): a statement expressing clear doubts about the impact is to be categorized as negative (-1). Examples: *"not sure what the impact is"*, *"the impact of the proposed studies is uncertain".* Statements in this direction have to be assessed carefully. If they describe that from the proposal it is not clear what the impact will be, this does not count as a statement expressing doubts and shall not be coded. Expressions indicating possibility (e.g. may, might, probably) do not indicate a negative valuation, as the authors of the evaluations can by definition not foresee with certainty the impact of the research.

61 clinical relevance / impact/improvements for patients / on treatment

Statements explaining how the project might benefit patients belong to this category. This fact can be phrased as changes in treatment, improvement in clinical practice guidelines etc. One has to be careful identifying statements referring to real improvements for patients. Guidelines, unless specifically pointed out as being policy relevant belong to this category. Examples:

+2 *"major impact on clinical practice", "the impact of this grant (...) is of great importance clinically for a better management of patients treatment.", "clinically highly relevant".*

+1 *"suggests a therapeutic chance for migraine patients", "could lead to better clinical post-stroke management of patients" "potential to revise guidelines for the perioperative management of patients".*

-1 *"unclear however whether the knowledge generated in this project could be effectively exploited for the therapy of type 1 diabetes"*

Examples of statements that are <u>not</u> to be coded: *"This is a clinically relevant questions"* → not clear whether the expert believes that it will have an impact or not. *"Clinical relevance"* → when there is no further specification it can not be coded. *"The proposal addresses a vital area in sphingolipid research with relevance to a human disease"* → it does not explicitly say that it will improve or have a positive impact on patients. *"The main strength of the proposal lay in its importance to the human health"* → it does not explicitly say that it is going to have an impact for patients. *"The concept will hopefully lead to clinical application in the future"* → due to the word hopefully it is not clear enough, the reviewer wishes for it to have an impact but doesn't say that it will have an impact. *"Proposal in a field of very high clinical importance"* → statement does not say something about the clinical importance of the study itself. *"Project that applies a technique to answer clinically significant questions"* → it does not explicitly say that it will improve or have a positive impact on patients.

62 useful for policy-/decision-makers

Statements about whether policy-/decision-makers might make use of the projects results. Requires some interpretation, as this fact might also be implicitly described. Examples: *"results will help interpreting post-mortem concentrations thus improving legal security", "policy significance".*

63 impact/advances in the academic field of study

Statements describing the projects' potential to develop approaches and methods in the field of study and beyond, its impact on the field. Example "*Project will have a significant impact in the field of toxicology".* A future change to the field should be described in the statements. Statements about how the proposed project is going to lead to (new) knowledge / a better understanding for the field belong to this category as well. Examples*: "improve our knowledge about the pathophysiology of these dieseases", "The studies will significantly improve our knowledge about this disease"*. Also statements about how the proposed project is going to provide (new) insights belong to this category. In this case one has to check, that the insights are in a way related to the field of study. Examples *"proposed project will develop insights about the role of correlation in population codes", "will provide new insight into islet cell plasticity"*. Statements referring to the generation of new data or having doubts whether it is going to lead to novel data are also to be included in this category. Example: *"This proposal is likely to lead to novel data"*. But *"produce interesting data"* is <u>not</u> coded. Statements that refer to the generation of new methods are also coded. Example: *"project will lead to new methods"*. But be careful because sometimes similar statements also just describe how new methods are already being developed at the moment. Important is always that it is explicit where and how the project is going to lead to new knowledge / insights / enhance understanding. Examples that are not sufficiently explicit: *"yield useful information", "results will be interesting", "project aims at understanding", "project is relevant for understanding xy*.

Statements that are <u>not</u> to be coded: being of interest for the field, example: *"The topic under investigation is of great interest at the level of basic science";* and being of importance for the field, examples: *"the main strength of the proposal lay in its importance to the sphingolipid world", "the problems to be tackled are of importance as models for the field", "hot topic", "relevant to the field", "focus on very important question"*.

The valuation in the context of this term needs careful assessment. Specific rules for the classification to the category of valuation:

+2 examples: *"if successful would represent a paradigm shift", "will have huge impact in forensic toxicology", "new era of understanding crop growth under stress"*

+1 examples: *"has the potential for being of wide use in the forensic field", "covers significant new ground in this area", "the approach could change the way the field looks for biomarkers of disease", "(...) opening opportunities beyond its own field", "contribution to the understanding of mechanism xy"*

-1 example: *"the project has limited posssibility to advance the knowledge of* *the discipline", "unclear how the acute exposure to GO will give insight into chronic disease states", "not clear what the project adds to the knowledge base"*

-2 example: *"project will not generate new knowledge", "this project does not have the potential to develop new approaches"*

70 exciting

This category does not require interpretation by the coder. Only the exact term (exciting) is coded. As a rule of thumb this category is only to be assigned to statements referring to the project as described in the proposal. It is <u>not</u> to be coded when it refers to predictions about impact or the generation of data. Examples: *"Project is likely to lead to novel and exciting data."* → code as 'novel' + 'positive' + 'general', *"Project opens new and exciting avenues for diabetes therapy"* → code as 'clinical relevance' + 'positive' + 'general. The classification to the category of valuation follows the standard rules.

80 state of the art

This category does not require interpretation by the coder. Only the exact term is coded. The classification to the category of valuation follows the standard rules. However it is anticipated that only the values +1 / -1 can be attributed to this category. If they refer to the literature and the state of the art described therein it is not going to be coded.

->introduced after confronting the scheme with the data. Not clear yet whether this concept really is related to innovation, because for example the European Patent Convention states that an invention is only considered new if it is not part of the state of the art. But maybe the researchers have different interpretation regarding this term.

90 cutting edge

The term requires some interpretation. Not only the term cutting edge, but also closely related terms (e.g. frontier technology, avant garde) are coded. The classification to the category of valuation follows the standard rules. However it is anticipated that only the values +1 / -1 can be attributed to this category.

-> *term was used in one of the guidelines*

100 risky

The term requires some interpretation. Not only the term risky/risk, but also closely related terms (e.g. speculative) are coded. The classification to the category of valuation follows the standard rules. High risk is positively valued whereas low risk is negative. Only rather general statements are to be assigned to this category. Examples: *"A high risk but potentially high reward project", "This is clearly a high-risk project", "High-risk project", "The project* *described is of very high-risk (and potentially of high-reward)".* Examples that shall not be coded: *"Risk that this method will not be suitable to work in neurons.", "This approach reduces the risk of the project to end up with no results.", "There is a risk that variations of adjuvant chemotherapy will affect the results of the study."*

99 Other, specify:

Other term used to describe innovation (OtherTerm)

Enter the term they used to describe innovation. Check first whether it cannot be added to a category already existing. E.g. pioneer could be added to the category "groundbreaking".

The following terms shall not be entered, because they are not considered being related to the concept of innovation: interesting, modern, up-to-date, need for more preliminary data, promising, brilliant, hot topic, fascinating, premature, challenging, enthusiasm/enthusiastic, preliminary.

Valuation (Valuation)

+2 very positive statement

Statements that include an adjective that strengthens the innovation aspect and gives it a more positive meaning are considered as very positive. Examples: *"very …", "highly …", "outstanding …", "huge impact", "scientific originality is excellent".* See also the explanations of the categories belonging to the variable *"term used to describe innovation"* for further specifications.

+1 positive statement

Neutral statements that do not include an adjective are considered positive. Examples: *"proposal is original", "proposal is innovative", "project is novel", "in part original".* See also the explanations of the categories belonging to the variable *"term used to describe innovation"* for further specifications.

-1 negative statement

Statements that include an adjective or another phrasing that weakens the innovation aspect and gives it a less positive meaning are considered negative. Examples: *"project is not fully original", "the project is little original", "not overwhelmingly/particularly original", "originality is modest", " a bit disappointed in terms of originality", "review methodology with a certain originality", "could have been more innovative", "does not go beyond that in innovativeness" "Results will come out but most likely no major breakthroughs" ,"can't imagine it will have broader impact".* See also the explanations of the categories belonging to the variable *"term used to describe innovation"* for further specifications.

-2 very negative statement

Statements that completely negate the innovation aspect are considered very negative. Examples: *"project is not original", "the project is not innovative, "in terms of innovation I am disappointed", "the idea is not novel", "strictly speaking the methodology is not original", "project not ground-breaking".* See also the explanations of the categories belonging to the variable *"term used to describe innovation"* for further specifications.

Aspect of application that is innovative (object)

What aspect of the application do they describe as being innovative? This component needs interpretation by the coder. First, because they do not always mention the exact term. Which means that interpretation (taking the context also into consideration) is warranted to allocate the description to the correct category. Example: *"Extending the project to ketones is innovative"* → From the background information one can deduce that ketones are the research material. Second, they might use an exact term, but describe something different or they use it in a meaning different to the one intended by this coding scheme. In this case the category described (and not the one explicitly mentioned) is to be selected. Example: *"The approach to use direct ketogenetic substances has never been tested in a RCT"* →By approach something bigger is meant in this coding scheme. The aspect he describes as being new is the research material / subject. Third, sometimes the aspect of application that is innovative might be explained in a different sentence than the declaration that something is innovative. Example: *"This is a very original idea. (...) The idea to use actual protein complexes instead of biomarker panels"* → shall be coded as research material / subject. Although these are two different sentences the explanation of what is actually innovative is coded together with the first sentence as one statement.

100 General

General statements about the project (or one or more subprojects) or its results as a whole or its expected results and with no further description belong to this category. Examples: *"project is very innovative", "proposal is original", "research is original", "the results may be used to inform policy".*

210 Topic of the project

Statements referring clearly to the topic / subject / theme of the project. The description of what is going to be examined. Examples: *"Topic of the project is original"*

220 Research aims / Research question

Statements about the goals / aims / objectives of the proposed project or its research question respectively the research problem. Example: "*Many of the questions have in some way already been addressed", "Specific research aims are not very original (...)"*

300 Theory / Hypotheses

Statements about the theory, theoretical concept, theoretical framework used in the application. Example: *"novel combination of known theories", "these two theories have not been combined so far", "theoretical concept is innovative"*.

Statements about the hypotheses. Example: *"novel hypothesis"*

400 Approach

Statements about the approach that will be employed in the proposed research. The approach is the way of considering / doing something or dealing with a situation / problem. Therefore also statements mentioning the idea / the concept (which is an abstract idea, plan or intention, but only if it is not further specified (e.g. theoretical concept would be categorized as theory) / the strategy / the design / the plan in the project / interventions are considered here. It is rather generic and long term. Example: *"using an experimental design is innovative", "The approach is innovative in that it combines both work conditioning and ergonomic approaches", "Innovative approaches in the field", "concept is original", "idea not entirely novel", "Scientific originality is excellent, given that double-blind, well-controlled studies are highly needed.", "approaches chosen for the proposed project are properly original"*.

410 Method

The method is more specific than approach (step by step, the manner in which work is executed, the practical realization of an approach, it is procedural). It is the particular way of doing something, so statements should be very specific to the proposed project. It is about how the more general approach has been adjusted to the specific project. Most of the statements use the term 'method' or 'methodology' directly, but also circumscriptions of these terms are attributed to this category. Example: *"method could be assessed as original", "review method with a certain originality"*

411 Research Material / Research Subject

Statements that specify what material / substances is going to be used / examined in the research project.

Example: *"The use of exogenous ketones represents an original and promising way to treat xy.", "There is originality in the proposed biomarkers"*

412 Technique / Technology

Statements that mention the technique / technology proposed by the applicants. Examples: *"The mass spectrometry techniques is state of the art", "project is mainly based on a novel technology", "the analytical method based on mass spectrometry with extended calibration range is also innovative".* If a technology / technique is described (without using the term technology / technique) a statement is also attributed to this category. Example: "the mode of sampling is innovative",

500 Unable to determine / other, specify

Select this category the statement can not be attributed to another aspect.

Other object (OtherObject)

Enter the object that has been described in the evaluation. If it could not have been attributed to one of the existing categories.

Annex 2. Reliability analysis

Coder reliability for the formal variables

IR_GradeTrackRecord	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

ER_GradeTrackRecord	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

IR_GradeProject	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

ER_GradeRelOrigTop	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

IR_CompRanking	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

ER_GradeMethFeas	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

ER_GradeOverall	
C_1 = 25	C_m = 24
C_2 = 25	C_R = 0.96

IR_Length[1]	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

ER_Length[1]	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

[1] For this variable a tolerance interval of 20 words was defined.

ER_OriginRecommendation	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

ER_Expertise	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

ER_Usefulness	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

ER_Count	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

ER_Origin	
C_1 = 25	C_m = 25
C_2 = 25	C_R = 1.00

IR_Quality	
C_1 = 25	C_m = 23
C_2 = 25	C_R = 0.92

ER_Quality	
C_1 = 25	C_m = 22
C_2 = 25	C_R = 0.88

IR_InnovationOverall (not used for analysis)	
C_1 = 25	C_m = 13
C_2 = 25	C_R = 0.52

ER_InnovationOverall (not used for analysis)	
C_1 = 25	C_m = 16
C_2 = 25	C_R = 0.64

Identification reliability for content variables

IR_Statements[2]	
C_1 = 39	C_m = 39
C_2 = 40	C_R = 0.987

ER_Statements[2]	
C_1 = 71	C_m = 69
C_2 = 72	C_R = 0.965

[2] When no statement was identified for an application this enters also into the number of codes.

Coder reliability for content variables

IR_Location	
C_1 = 28	C_m = 28
C_2 = 28	C_R = 1.00

IR_Location	
C_1 = 65	C_m = 65
C_2 = 65	C_R = 1.00

IR_Term	
C_1 = 28	C_m = 28
C_2 = 28	C_R = 1.00

IR_Term	
C_1 = 65	C_m = 62
C_2 = 65	C_R = 0.954

IR_Valuation	
C_1 = 28	C_m = 27
C_2 = 28	C_R = 0.964

IR_Valuation	
C_1 = 65	C_m = 62
C_2 = 65	C_R = 0.954

IR_Location	
C_1 = 28	C_m = 26
C_2 = 28	C_R = 0.929

IR_Object	
C_1 = 65	C_m = 59
C_2 = 65	C_R = 0.908

Explanation of the calculation

Name of Variable

C_1 = Number of codes in analysis 1	C_m = Number of matching codes
C_2 = Number of codes in analysis 2	C_R = Reliability coefficient
Holsti-formula : $C_R = 2 * C_m / (C_1 + C_2)$	

Annex 3. Tables and figures for descriptive analysis

Comparison of the samples with the population

Table 5.1. Funding status of applications (Sample 1)

	n_w	$\%_w$	N	%
Approved	51	45.0	429	46.5
Rejected	62	55.0	493	53.5

Table 5.2. Final rating of applications by the NRC (Sample 1)

	n_w	$\%_w$	N	%
A	5	4.9	25	2.7
AB	10	9.1	159	17.2
B	35	31.2	323	35.0
BC	36	32.3	217	23.5
C	19	16.5	152	16.5
D	7	6.1	44	4.8
Unable to determine	0	0.0	2	0.2

Table 5.3. Discipline group of applications (Sample 1)

	n_w	$\%_w$	N	%
Basic Biological Research	20	18.0	182	19.7
General Biology	9	7.9	94	10.2
Basic Medical Sciences	22	19.2	142	15.4
Experimental Medicine	23	20.1	151	16.4
Clinical Medicine	24	20.8	273	29.6
Preventive Medicine	15	13.5	60	6.5
Social Medicine	1	0.6	20	2.2

Table 5.4. Applicants' gender (Sample 1)

	n_w	$\%_w$	N	%
Female	28	25.2	199	21.6
Male	85	74.8	723	78.4

Table 5.5. Applicants' academic degree (Sample 1)

	n_w	$\%_w$	N	%
No degree	0	0.0	3	0.3
Doctoral degree	61	53.6	429	46.5
Professor	52	46.4	490	53.1

Table 5.6. Applicants' age (Sample 1)

	n_w	$\%_w$	N	%
-35 yrs	10	9.2	65	7.0
36-40 yrs	21	18.4	184	20.0
41-45 yrs	42	37.2	249	27.0
46-50 yrs	17	15.2	202	21.9
51-55 yrs	12	10.4	126	13.7
56-60 yrs	4	3.9	71	7.7
60+ yrs	6	5.7	25	2.7

Table 5.7. Institutional affiliation of applications (Sample 2)

	n_w	$\%_w$	N	%
ETH domain	17	28.9	62	19.2
Cantonal universities	39	65.3	228	70.6
UAS	1	2.4	7	2.2
Others	2	3.3	26	8.0

Table 5.8. Discipline group of applications (Sample 2)

	n_w	$\%_w$	N	%
Basic Biological Research	15	24.7	73	22.6
General Biology	6	10.7	38	11.8
Basic Medical Sciences	11	18.2	55	17.0
Experimental Medicine	8	13.6	46	14.2
Clinical Medicine	18	30.2	93	28.8
Preventive Medicine	1	0.9	14	4.3
Social Medicine	1	1.8	4	1.2

Table 5.9. Applicants' gender (Sample 2)

	n_w	$\%_w$	N	%
Female	10	16.7	68	21.1
Male	50	83.3	255	78.9

Table 5.10. Applicants' academic degree (Sample 2)

	n_w	$\%_w$	N	%
No degree	0	0.0	0	0.0
Doctoral degree	23	38.0	164	50.8
Professor	37	62.0	159	49.2

Table 5.11. Applicants' age (Sample 2)

	n_w	$\%_w$	N	%
-35 yrs	6	10.7	29	9.0
36-40 yrs	14	23.1	68	21.1
41-45 yrs	19	31.6	93	28.8
46-50 yrs	11	19.1	64	19.8
51-55 yrs	7	12.0	36	11.1
56-60 yrs	1	0.9	26	8.0
60+ yrs	2	2.7	7	2.2

Characteristics of the reviews

Table 5.12. Length of external reviews (Sample 1)

	n_w	$\%_w$
-250 words	10	8.8
251-500 words	39	34.6
501-750 words	24	21.0
751-1000 words	22	19.0
1001-1250 words	9	8.3
1250+ words	9	8.3

Table 5.13. Quality of external reviews (Sample 1)

	n_w	$\%_w$
Bad	15	13.6
Good	98	86.4

Table 5.14. Source of recommendation of external reviewer (Sample 1)

	n_w	$\%_w$
Recommended by applicant	15	13.3
Chosen by the SNSF	98	86.7

Table 5.15. Usefulness of external reviews (Sample 1)

	n_w	$\%_w$
Useful	95	83.6
In part useful	18	16.4

Table 5.16. Number of external reviews per application (Sample 1)

	$\bar{x}_w$	μ
Reviews / Application	2.5	2.9

Table 5.17. Country of residence of external reviewer (Sample 1)

	n_w	$\%_w$
Switzerland	42	37.3
Other, not Switzerland	71	62.7

Table 5.18. Expertise of external reviewer (Sample 1)

	n_w	$\%_w$
is within my area of specialisation	66	58.8
is withn my wider discipline	47	41.2

Table 5.19. Length of recommendations (Sample 1)

	n_w	$\%_w$
-250 words	34	30.1
251-500 words	55	48.2
501-750 words	18	16.4
751-1000 words	4	3.6
1001-1250 words	2	1.5
1250+ words	0	0.2

Table 5.20. Quality of recommendations (Sample 1)

	n_w	$\%_w$
Bad	1	1.2
Good	112	98.8

Table 5.21. Average length of review

	$\bar{x}_w$
External review (Sample 1)	666
Recommendation (Sample 1)	384
Recommendation (Sample 2)	410

Table 5.22. Length of recommendations (Sample 2)

	n_w	$\%_w$
-250 words	22	36.0
251-500 words	24	39.6
501-750 words	7	11.8
751-1000 words	6	9.3
1001-1250 words	1	0.9
1250+ words	1	2.4

Table 5.23. Quality of recommendations (Sample 2)

	n_w	$\%_w$
Bad	4	6.7
Good	56	93.3

Analysis of the identified statements

Table 5.24. Number of statements identified

	n	n_w	$\bar{x}_w$
External Reviews (Sample 1)	330	316	2.8
Recommendations (Sample 1)	175	139	1.2
Recommendations (Sample 2)	98	91	1.5

Table 5.25. Statements in external reviews (Sample 1)

	n_w	$\%_w$
no statement at all	30	26.9
1 statement	12	10.6
2 statements	9	8.1
3 statements	22	19.9
4 statements	7	6.3
5 statements	20	17.3
6 statements	2	1.5
7 statements	3	2.7
8 statements	7	6.1
9 statements	1	0.5

Table 5.26. Location of statements in external reviews (Sample 1)

	n_w	$\%_w$
Comments regarding the overall assessment	88	28.0
Scientific relevance, orginality and topicality	200	63.1
Suitability of methods and feasibility	28	8.9

Table 5.27. Valuation of statements in external reviews (Sample 1)

	n_w	$\%_w$
Very negative statement	11	3.4
Negative statement	19	6.1
Positive statement	245	77.5
Very positive statement	41	13.0

Table 5.28. Statements mention of aspect of application in external reviews (Sample 1)

	n_w	$\%_w$
General	187	59.3
Topic of the project	24	7.7
Research aims	8	2.4
Theory / Hypotheses	0	0.0
Approach	65	20.7
Method	17	5.4
Research Material / Subject	3	1.0
Technique/Technology	10	3.3
Unable to determine / Other	1	0.3

Table 5.29. Statements in recommendations (Sample 1)

	n_w	$\%_w$
no statement at all	47	41.4
1 statement	27	24.2
2 statements	24	20.8
3 statements	6	5.2
4 statements	4	3.9
5 statements	1	1.0
6 statements	4	3.2
7 statements	0	0.2

Table 5.30. Location of statements in recommendations (Sample 1)

	n_w	$\%_w$
Assessment of the proposed project	91	64.6
Comments regarding comparative ranking	48	35.4

Table 5.31. Valuation of statements in recommendations (Sample 1)

	n_w	$\%_w$
Very negative statement	9	6.1
Negative statement	38	27.2
Positive statement	84	60.3
Very positive statement	9	6.4

Table 5.32. Statements mention of aspect of application in recommendations (Sample 1)

	n_w	$\%_w$
General	110	79.0
Topic of the project	0	0.2
Research aims	10	7.1
Theory / Hypotheses	0	0.0
Approach	11	7.7
Method	7	4.7
Research Material / Subject	0	0.0
Technique/Technology	1	1.1
Unable to determine / other	0	0.2

Table 5.33. Statements in recommendations (Sample 2)

	n_w	$\%_w$
no statement at all	15	25.8
1 statement	18	29.8
2 statements	15	25.1
3 statements	6	10.0
4 statements	4	6.7
5 statements	2	2.7

Table 5.34. Location of statements in recommendations (Sample 2)

	n_w	$\%_w$
Assessment of the proposed project	46	51.0
Comments regarding comparative ranking	45	49.0

Table 5.35. Valuation of statements in recommendations (Sample 2)

	n_w	$\%_w$
Very negative statement	2	2.3
Negative statement	20	21.9
Positive statement	61	66.6
Very positive statement	8	9.2

Table 5.36. Statements mention of aspect of application in recommendations (Sample 2)

	n_w	$\%_w$
General	38	41.4
Topic of the project	6	7.0
Research aims	19	20.6
Theory / Hyotheses	1	1.6
Approach	18	19.2
Method	4	4.7
Research Material / Subject	0	0.0
Technique/Technology	5	5.5
Unable to determine / not known	0	0.0

Table 5.37. Statistics about indices (Sample 1: External reviews)

	Index “perceived degree of innovation”	Index “anticipated impact"
$\overline{x}_w$	0.935	0.763
$\tilde{x}_w$	1	1
σ_w	1.013	1.063
$x_{w\ min}$	-3.500	-1.000
$x_{w\ max}$	3.667	5.000

Table 5.38. Innovative aspect of the application (Sample 1: External reviews)

	"project in general"		"research goal / topic"		"theory"		"approach / method"	
	n_w	$\%_w$	n_w	$\%_w$	n_w	$\%_w$	n_w	$\%_w$
Value is negative	4	3.5	1	0.8	0	0.0	8	7.3
Value is 0	61	54.0	91	80.9	113	100.0	71	62.7
Value is between 0 and 2	39	34.7	20	18.1	0	0.0	33	29.4
Value equals 2 or greater	9	7.8	0	0.2	0	0.0	1	0.7

Table 5.39. Types of impcat (Sample 1: External reviews)

	"impact on patient / treatment"		"impact / advances in the academic field"		"other impact"	
	n_w	$\%_w$	n_w	$\%_w$	n_w	$\%_w$
Value is negative	0	0.1	4	3.2	0	0.2
Value is 0	96	85.1	59	52.3	103	91.0
Value is between 0 and 2	14	12.3	44	39.2	6	5.7
Value equals 2 or greater	3	2.5	6	5.3	3	3.0

Table 5.40. Statistics about indices (Sample 1: Recommendations)

	Index "perceived degree of innovation"	Index "anticipated impact"
$\bar{x}_w$	0.017	0.233
$\tilde{x}_w$	0	0
σ_w	0.995	0.787
$x_{w\ min}$	-5.000	-2.000
$x_{w\ max}$	2.500	3.000

Table 5.41. Innovative aspect of the application (Sample 1: Recommendations)

	"project in general"		"research goal / topic"		"theory"		"approach / method"	
	n_w	$\%_w$	n_w	$\%_w$	n_w	$\%_w$	n_w	$\%_w$
Value is negative	20	18.1	4	3.4	0	0.0	5	4.3
Value is 0	75	66.4	105	93.1	113	100.0	98	86.9
Value is between 0 and 2	17	14.8	4	3.6	0	0.0	10	8.8
Value equals 2 or greater	1	0.7	0	0.0	0	0.0	0	0.0

Table 5.42. Types of impcat (Sample 1: Recommendations)

	"impact on patient / treatment"		"impact / advances in the academic field"		"other impact"	
	n_w	$\%_w$	n_w	$\%_w$	n_w	$\%_w$
Value is negative	0	0.2	4	3.3	0	0.0
Value is 0	103	91.1	91	80.8	108	95.4
Value is between 0 and 2	10	8.6	18	15.9	5	4.6
Value equals 2 or greater	0	0.1	0	0.0	0	0.0

Table 5.43. Statistics about indices (Sample 2)

	Index "perceived degree of innovation"	Index "anticipated impact"
$\bar{x}_w$	0.690	0.073
$\tilde{x}_w$	1	0
σ_w	1.154	0.442
$x_{w\ min}$	-2.000	-2.000
$x_{w\ max}$	3.000	1.000

Table 5.44. Innovative aspect of the application (Sample 2)

	"project in general"		"research goal / topic"		"theory"		"approach / method"	
	n_w	$\%_w$	n_w	$\%_w$	n_w	$\%_w$	n_w	$\%_w$
Value is negative	7	12.0	0	0.0	0	0.0	2	2.7
Value is 0	39	65.6	45	75.3	59	97.6	43	71.6
Value is between 0 and 2	10	16.7	15	24.7	1	2.4	13	22.4
Value equals 2 or greater	3	5.8	0	0.0	0	0.0	2	3.3

Table 5.45. Types of impcat (Sample 2)

	"impact on patient / treatment"		"impact / advances in the academic field"		"other impact"	
	n_w	$\%_w$	n_w	$\%_w$	n_w	$\%_w$
Value is negative	0	0.2	2	2.7	1	1.8
Value is 0	57	95.8	55	92.2	57	94.9
Value is between 0 and 2	3	4.2	3	5.1	2	3.3
Value equals 2 or greater	0	0.1	0	0.0	0	0.0

Annex 4. Tables and figures for inferential statistical analysis

Table 6.1. Test for multicollinearity (Model 1a)

	Collinearity Statistics	
	Tolerance	VIF
Index "perceived degree of innovation"	0.733	1.365
Index "anticipated impact"	0.737	1.357
Applicant's gender	0.825	1.211
Applicant's age	0.766	1.305
Applicant's academic degree	0.752	1.331
Discipline group	0.767	1.304
Type of institution	0.931	1.075
Country of residence of external reviewer	0.750	1.333
Source of recommendation	0.830	1.204
Expertise of external reviewer	0.857	1.167

Dependent Variable: Grade overall assessment

Table 6.2. Test for multicollinearity (Model 1.1a)

	Collinearity Statistics	
	Tolerance	VIF
Aspect "project in general"	0.614	1.629
Aspect "research goal / topic"	0.788	1.269
Aspect "approach / method"	0.784	1.275
Impact type "impact on patient / treatment"	0.476	2.099
Impact type "impact / advances in the academic field"	0.644	1.552
Impact type "other impact"	0.589	1.698
Applicant's gender	0.664	1.505
Applicant's age	0.648	1.543
Applicant's academic degree	0.710	1.408
Discipline group	0.679	1.473
Type of institution	0.893	1.120
Country of residence of external reviewer	0.756	1.322
Source of recommendation	0.747	1.339
Expertise of external reviewer	0.734	1.363

Dependent Variable: Grade overall assessment

Table 6.3. Test for multicollinearity (Model 1b)

	Collinearity Statistics	
	Tolerance	VIF
Index "perceived degree of innovation"	0.892	1.121
Index "anticipated impact"	0.920	1.087
Applicant's gender	0.911	1.098
Applicant's age	0.791	1.264
Applicant's academic degree	0.791	1.265
Discipline group	0.794	1.259
Type of institution	0.940	1.064

Dependent Variable: Grade comparative ranking

Table 6.4. Test of parallel lines[a] (Model 1a)

	-2 Log-Likelihood	Chi-Quadrat	df	Sig.
Null Hypothesis	204.334			
General	133.301[b]	71.032[c]	84	0.843

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

b. The log-likelihood value cannot be further increased after maximum number of step-halving.

c. The Chi-Square statistic is computed based on the log-likelihood value of the last iteration of the general model. Validity of the test is uncertain.

Table 6.5. Test of parallel lines[a] (Model 1.1a)

	-2 Log-Likelihood	Chi-Quadrat	df	Sig.
Null Hypothesis	186.675			
General	80.822[b]	105.852[c]	100	0.325

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

b. The log-likelihood value cannot be further increased after maximum number of step-halving.

c. The Chi-Square statistic is computed based on the log-likelihood value of the last iteration of the general model. Validity of the test is uncertain.

Table 6.6. Test of parallel lines[a] (Model 1b)

	-2 Log-Likelihood	Chi-Quadrat	df	Sig.
Null Hypothesis	237.611			
General	.000[b]	237.611	72	0.000

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

b. The log-likelihood value is practically zero. There may be a complete separation in the data. The maximum likelihood estimates do not exist.

Table 6.7. Confusion table (Model 1a)

			Predicted grade overall assessment						
			poor	average	good	very good	excellent	out-standing	Total
Observed grade overall assessment	poor	Count	3	0	0	0	0	0	3
		% within observed grade overall assessment	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
	average	Count	0	6	0	1	0	0	7
		% within observed grade overall assessment	0.0%	85.7%	0.0%	14.3%	0.0%	0.0%	100.0%
	good	Count	0	3	10	6	0	0	19
		% within observed grade overall assessment	0.0%	15.8%	52.6%	31.6%	0.0%	0.0%	100.0%
	very good	Count	0	0	1	30	5	1	37
		% within observed grade overall assessment	0.0%	0.0%	2.7%	81.1%	13.5%	2.7%	100.0%
	excellent	Count	0	1	1	6	18	3	29
		% within observed grade overall assessment	0.0%	3.4%	3.4%	20.7%	62.1%	10.3%	100.0%
	out-standing	Count	0	0	0	0	5	12	17
		% within observed grade overall assessment	0.0%	0.0%	0.0%	0.0%	29.4%	70.6%	100.0%
Total		Count	3	10	12	43	28	16	112
		% within observed grade overall assessment	2.7%	8.9%	10.7%	38.4%	25.0%	14.3%	100.0%

Table 6.8. Confusion table (Model 1.1a)

			Predicted grade overall assessment						
			poor	average	good	very good	excellent	out-standing	Total
Observed grade overall assessment	poor	Count	3	0	0	0	0	0	3
		% within observed grade overall assessment	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
	average	Count	0	4	3	1	0	0	8
		% within observed grade overall assessment	0.0%	50.0%	37.5%	12.5%	0.0%	0.0%	100.0%
	good	Count	0	0	18	0	0	0	18
		% within observed grade overall assessment	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%
	very good	Count	1	0	0	32	3	1	37
		% within observed grade overall assessment	2.7%	0.0%	0.0%	86.5%	8.1%	2.7%	100.0%
	excellent	Count	0	0	2	3	22	3	30
		% within observed grade overall assessment	0.0%	0.0%	6.7%	10.0%	73.3%	10.0%	100.0%
	out-standing	Count	0	0	0	0	2	15	17
		% within observed grade overall assessment	0.0%	0.0%	0.0%	0.0%	11.8%	88.2%	100.0%
Total		Count	4	4	23	36	27	19	113
		% within observed grade overall assessment	3.5%	3.5%	20.4%	31.9%	23.9%	16.8%	100.0%

Table 6.9. Test of model effects (Model 1a)

	Wald Chi-Square	df	Sig.
Index "perceived degree of innovation"	0.547	1	0.460
Index "anticipated impact"	48.961	1	0.000
Applicant's gender	29.571	1	0.000
Applicant's age	9.925	5	0.077
Applicant's academic degree	0.057	1	0.812
Discipline group	48.935	6	0.000
Type of institution	4.164	3	0.244
Country of residence of external reviewer	0.130	1	0.718
Source of recommendation	37.946	1	0.000
Expertise of external reviewer	6.246	1	0.012

Dependent Variable: Grade overall assessment

Table 6.10. Test of model effects (Model 1.1a)

	Wald Chi-Square	df	Sig.
Aspect "project in general"	0.460	1	0.498
Aspect "research goal / topic"	9.038	1	0.003
Aspect "approach / method"	8.155	1	0.004
Impact type "impact on patient / treatment"	13.213	1	0.000
Impact type "impact / advances in the academic field"	17.316	1	0.000
Impact type "other impact"	24.790	1	0.000
Applicant's gender	32.427	1	0.000
Applicant's age	11.011	5	0.051
Applicant's academic degree	0.346	1	0.556
Discipline group	53.357	6	0.000
Type of institution	5.973	3	0.113
Country of residence of external reviewer	0.426	1	0.514
Source of recommendation	35.516	1	0.000
Expertise of external reviewer	0.022	1	0.882

Dependent Variable: Grade overall assessment

Table 6.11. Parameter estimates (Model 1a)

		Sig.	Odds Ratio	95% Confidence Interval	
				Lower	Upper
Threshold	Grade overall assessment = poor	0.000	0.000	0.000	0.000
	Grade overall assessment = average	0.000	0.000	0.000	0.000
	Grade overall assessment = good	0.000	0.000	0.000	0.005
	Grade overall assessment = very good	0.001	0.002	0.000	0.091
	Grade overall assessment = excellent	0.188	0.099	0.003	3.096
Index "perceived degree of innovation"		0.460	0.837	0.523	1.341
Index "anticipated impact"		0.000	11.486	5.797	22.759
Applicant's gender: female		0.000	0.032	0.009	0.111
Applicant's gender: male			1		
Applicant's age: 55+ yrs		0.670	1.616	0.177	14.747
Applicant's age: 51-55 yrs		0.018	12.450	1.541	100.590
Applicant's age: 46-50 yrs		0.438	2.078	0.327	13.218
Applicant's age: 41-45 yrs		0.600	1.531	0.312	7.506
Applicant's age: 36-40 yrs		0.706	0.715	0.125	4.099
Applicant's age: -35 yrs			1		
Applicant's academic degree: PhD		0.812	0.888	0.333	2.365
Applicant's academic degree: Professor			1		
Discipline group: 309		0.625	3.851	0.017	860.955
Discipline group: 308		0.000	0.007	0.001	0.051
Discipline group: 307		0.395	2.015	0.402	10.107
Discipline group: 304		0.000	0.018	0.004	0.088
Discipline group: 303		0.032	0.200	0.046	0.872
Discipline group: 302		0.004	0.053	0.007	0.389
Discipline group: 301			1		
Type of institution: Others		0.216	0.152	0.008	3.001
Type of institution: Cantonal universities		0.776	0.688	0.053	8.984
Type of institution: ETH domain		0.936	1.120	0.069	18.066
Type of institution: Universities of applied sciences			1.000		
Country of residence of external reviewer: else		0.718	0.842	0.330	2.146
Country of residence of external reviewer: Switzerland			1		
Source of recommendation: chosen by SNSF		0.000	0.002	0.000	0.017
Source of recommendation: applicant's list			1		
Expertise of external reviewer: within wider discipline		0.012	0.295	0.113	0.769
Expertise of external reviewer: within area of specialisation			1		

Dependent Variable: Grade overall assessment

Model fit	
Nagelkerke's Pseudo R^2	0.778
Likelihood-ratio test	0.000

Table 6.12. Parameter estimates (Model 1.1a)

		Sig.	Odds Ratio	95% Confidence Interval	
				Lower	Upper
Threshold	Grade overall assessment = poor	0.000	0.000	0.000	0.000
	Grade overall assessment = average	0.000	0.000	0.000	0.001
	Grade overall assessment = good	0.000	0.000	0.000	0.007
	Grade overall assessment = very good	0.006	0.003	0.000	0.195
	Grade overall assessment = excellent	0.620	0.394	0.010	15.614
Aspect "project in general"		0.498	1.302	0.607	2.796
Aspect "research goal / topic"		0.003	5.515	1.812	16.789
Aspect "approach / method"		0.004	0.241	0.091	0.640
Impact type "impact on patient / treatment"		0.000	19.538	3.934	97.034
Impact type "impact / advances in the academic field"		0.000	7.796	2.963	20.508
Impact type "other impact"		0.000	265.816	29.522	2393.430
Applicant's gender: female		0.000	0.011	0.002	0.051
Applicant's gender: male			1		
Applicant's age: 55+ yrs		0.503	2.342	0.194	28.196
Applicant's age: 51-55 yrs		0.010	25.013	2.125	294.456
Applicant's age: 46-50 yrs		0.534	1.913	0.247	14.815
Applicant's age: 41-45 yrs		0.099	4.405	0.755	25.691
Applicant's age: 36-40 yrs		0.724	1.411	0.209	9.519
Applicant's age: -35 yrs			1		
Applicant's academic degree: PhD		0.556	0.733	0.260	2.066
Applicant's academic degree: Professor			1		
Discipline group: 309		0.586	4.632	0.019	1154.184
Discipline group: 308		0.000	0.001	0.000	0.013
Discipline group: 307		0.175	3.387	0.581	19.740
Discipline group: 304		0.000	0.018	0.003	0.094
Discipline group: 303		0.146	0.295	0.057	1.532
Discipline group: 302		0.092	0.164	0.020	1.347
Discipline group: 301			1.000		
Type of institution: Others		0.124	0.089	0.004	1.950
Type of institution: Cantonal universities		0.324	0.260	0.018	3.779
Type of institution: ETH domain		0.978	0.960	0.054	17.136
Type of institution: Universities of applied sciences			1		
Country of residence of external reviewer: else		0.514	1.399	0.511	3.830
Country of residence of external reviewer: Switzerland			1		
Source of recommendation: chosen by SNSF		0.000	0.002	0.000	0.015
Source of recommendation: applicant's list			1		
Expertise of external reviewer: within wider discipline		0.882	0.920	0.309	2.742
Expertise of external reviewer: within area of specialisation			1		

Dependent Variable: Grade overall assessment

Model fit	
Nagelkerke's Pseudo R^2	0.816
Likelihood-ratio test	0.000

Table 6.13. Confustion table (Model 1b)

			Predicted grade comparative ranking						Total
			D	C	BC	B	AB	A	
Observed grade comparative ranking	D	Count	0	3	0	0	0	0	3
		% within observed grade overall assessment	0.0%	100.0 %	0.0%	0.0%	0.0%	0.0%	100.0%
	C	Count	0	12	5	4	0	0	21
		% within observed grade overall assessment	0.0%	57.1%	23.8%	19.0%	0.0%	0.0%	100.0%
	BC	Count	0	6	20	11	0	0	37
		% within observed grade overall assessment	0.0%	16.2%	54.1%	29.7%	0.0%	0.0%	100.0%
	B	Count	0	0	7	27	2	0	36
		% within observed grade overall assessment	0.0%	0.0%	19.4%	75.0%	8.1%	0.0%	100.0%
	AB	Count	0	0	0	8	0	3	11
		% within observed grade overall assessment	0.0%	0.0%	0.0%	72.7%	0.0%	27.3%	100.0%
	A	Count	0	0	0	5	0	0	5
		% within observed grade overall assessment	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	100.0%
Total		Count	0	21	32	55	2	3	113
		% within observed grade overall assessment	0.0%	18.6%	28.3%	48.7%	1.8%	2.7%	100.0%

Table 6.14. Test of model effects (Model 1b)

	Wald Chi-Square	df	Sig.
Index "perceived degree of innovation"	13.190	1	0.000
Index "anticipated impact"	10.782	1	0.001
Applicant's gender	0.679	1	0.410
Applicant's age	18.260	5	0.003
Applicant's academic degree	6.665	1	0.010
Discipline group	37.772	6	0.000
Type of institution	1.589	3	0.662

Dependent Variable: Grade comparative ranking

Table 6.15. Parameter estimates (Model 1b)

	Sig.	Odds Ratio	95% Confidence Interval Lower	Upper
Threshold Grade comparative ranking = D	0	0.001	0	0.025
Grade comparative ranking = C	0.016	0.022	0.001	0.496
Grade comparative ranking = BC	0.432	0.29	0.013	6.354
Grade comparative ranking = BC	0.267	5.521	0.27	112.846
Grade comparative ranking = AB	0.025	34.838	1.572	772.235
Index "perceived degree of innovation"	0.000	2.208	1.440	3.386
Index "anticipated impact"	0.001	2.581	1.465	4.544
Applicant's gender: female	0.410	1.479	0.583	3.748
Applicant's gender: male		1		
Applicant's age: 55+ yrs	0.153	4.121	0.592	28.680
Applicant's age: 51-55 yrs	0.608	1.632	0.252	10.578
Applicant's age: 46-50 yrs	0.347	0.453	0.087	2.359
Applicant's age: 41-45 yrs	0.745	1.268	0.304	5.284
Applicant's age: 36-40 yrs	0.028	0.173	0.037	0.824
Applicant's age: -35 yrs		1		
Applicant's academic degree: PhD	0.010	0.301	0.121	0.749
Applicant's academic degree: Professor		1		
Discipline group: 309	0.203	0.028	0.000	6.971
Discipline group: 308	0.000	0.005	0.001	0.031
Discipline group: 307	0.265	0.441	0.104	1.861
Discipline group: 304	0.205	0.418	0.109	1.608
Discipline group: 303	0.266	0.453	0.112	1.827
Discipline group: 302	0.146	0.274	0.048	1.572
Discipline group: 301		1		
Type of institution: Others	0.553	2.329	0.143	38.034
Type of institution: Cantonal universities	0.871	1.220	0.110	13.521
Type of institution: ETH domain	0.519	2.344	0.176	31.192
Type of institution: Universities of applied sciences		1.000		

Dependent Variable: Grade overall assessment

Model fit	
Nagelkerke's Pseudo R^2	0.610
Likelihood-ratio test	0.000

Table 6.16. Test for multicollinearity (Model 2)

	Collinearity Statistics	
	Tolerance	VIF
Index "perceived degree of innovation"	0.980	1.020
Index "anticipated impact"	0.930	1.075
Applicant's gender	0.957	1.045
Applicant's age	0.935	1.069
Applicant's academic degree	0.855	1.169
Year of ruling	0.902	1.109

Dependent Variable: Funding status

L’IDHEAP en un coup d’oeil

Champ

L’IDHEAP, créé en 1981, se concentre sur l’étude de l’administration publique, un champ interdisciplinaire (en anglais Public Administration) visant à développer les connaissances scientifiques sur la conduite des affaires publiques et la direction des institutions qui en sont responsables. Ces connaissances s’appuient sur plusieurs disciplines des sciences humaines et sociales, comme le droit, l’économie, le management et la science politique, adaptées aux spécificités du secteur public et parapublic. L’IDHEAP est le seul institut universitaire suisse totalement dédié à cet important champ de la connaissance. Il est intégré dans la Faculté de droit, des sciences criminelles et d’administration publique de l’Université de Lausanne.

Vision

A l’interface entre théorie et pratique de l’administration publique, l’IDHEAP est le pôle national d’excellence contribuant à l’analyse des mutations du secteur public et à une meilleure gouvernance de l’Etat de droit à tous ses niveaux, en pleine coopération avec ses partenaires universitaires suisses et étrangers.

Mission

Au service de ses étudiants, du secteur public et de la société dans son ensemble, l’IDHEAP a une triple mission qui résulte de sa vision:

- Enseignement universitaire accrédité au niveau master et post-master, ainsi que formation continue de qualité des élus et cadres publics;
- Recherche fondamentale et appliquée en administration publique reconnue au niveau national et international, et valorisée dans le secteur public suisse;
- Expertise et conseil indépendants appréciés par les organismes publics mandataires et enrichissant l’enseignement et la recherche.