Vetenskapsrådet

# QUALITY ASSESSMENT IN PEER REVIEW

## A Swedish Research Council seminar

VETENSKAPSRÅDETS LILLA RAPPORTSERIE 1:2011

# QUALITY ASSESSMENT IN PEER REVIEW

## A Swedish Research Council seminar

Stockholm, 5 November 2009

QUALITY ASSESSMENT IN PEER REVIEW

VETENSKAPSRÅDET
Box 1035
101 38 Stockholm


ISBN 978-91-7307-190-1

# PREFACE

Discussions about quality and rating of quality parameters are frequent among organisations that fund research, e.g. the Swedish Research Council (SRC). Evaluation of research proposals, i.e. assessments as a tool to review research quality, is one of several aspects. In this context, grading scales in general, as well as the grading system used by the SRC more specifically, are discussed with regard refinement of the evaluation of proposals.

Another aspect is how grading systems can be used to compare quality within or between research areas. Funders need an instrument to compare the quality of scientific research in order to distinguish research with the highest quality and to prioritise projects. There are also reasons to use well-developed evaluation criteria and grading scales in communication concerning research and research funding.

On 5 November 2009, the Swedish Research Council hosted a workshop on grading scales in the evaluation process with focus on peer review. The presentations during the day gave an international perspective of other countries' research councils and their use of grading scales when evaluating applications for research funding. Experts on grading systems and grading scales gave a scientific introduction to the discussions. This report, written by Anna Lagerkvist, is a summary of the discussions at the conference. The aim is to provide support for the continuous work with evaluations within the Swedish Research Council.

The mission of the Swedish Research Council is to support Swedish research of the highest quality. To maintain and develop the core values of peer review, it is important to continuously monitor the process by which the quality assessments are made. The workshop on grading scales is one such effort.

Stockholm in June 2010,

**Karin Forsberg Nilsson**
*former Deputy Secretary General*
*Swedish Research Council, Medicine and Health*

# INNEHÅLL

# INTRODUCTION

The seminar took place at Vetenskapsrådet (Swedish Research Council) in Stockholm on 5 November 2009. Participants included representatives from funding organisations, universities, research councils and other stake holders interested in the quality of peer review. Professor Karin Forsberg Nilsson, Deputy Secretary General, Swedish Research Council Medicine, welcomed the participants. The seminar, served also as a starting point to an intensified discussion within Swedish Research Council (SRC) on grading scales in peer review.

"Funding agencies don't always ask themselves why they use a particular scale in the grading process. We would like to share experience between research councils and also receive scientific guidance. That is the aim of this seminar," Forsberg Nilsson said.

# SESSION 1: SHARING THE EXPERIENCE OF FUNDING ORGANISATIONS ON SCORING OF APPLICATIONS

This session aimed at showcasing the different research environments to date in countries such as Sweden, Norway, Finland and the US. There are huge differences between the various nations, the audience heard.

## Sweden

Professor Håkan Billig, Secretary General, Medicine at the Swedish Research Council (SRC) started the seminar by giving an overview of peer review processes in Sweden. Some SEK 110 billion (€10.6 billion) is spent on research and development in Sweden every year. The Swedish Research Council funds basic research of the highest quality in all areas. It receives some 5,000 applications each year. Its budget for 2009 was SEK 4 billion.

Each proposal is evaluated and funded by its associated scientific council or committee. These include Humanities and Social Sciences, Medicine[1], Natural and Engineering Sciences, Educational Science, and Research Infrastructure. Some 500 experts take part in the evaluation panels, appointed by the research councils or committees.

"The peer review process is key for the Swedish Research Council," said Håkan Billig. "This system is central to our operation."

The research bill, introduced in 2008, directs more of the funds to be distributed to universities, some directly and some based on quality indicators. The bill also earmarks a total of SEK 1.8 billion for areas specified by the government to be of strategic importance for Swedish research.

"The introduction of the new research bill was a dramatic change for Swedish research funding. In Europe, it is very unusual for governments to make detailed funding decisions on specific projects itself," Billig commented.

During peer review within the SRC, panels of experts review the applications. Depending on the council, the applications are reviewed by five experts in the panel itself (Medicine) or reviewed by fewer panel members and sent on to external reviewers. These provide additional reviews on a few applications each (Natural and Engineering Sciences). Common to all councils and committees is that the panels are set up to match specific areas of research. After individual assessment of the applications, panels meet to discuss the proposals and make recommendations for funding to the councils and committees.

The Scientific Council for Medicine panels are organised according to disease area and flexible in number to match the amount of applications submitted. This organisation has recently been evaluated by an external expert panel.[2] Each application is read by a minimum of five peers, who grade the applications individually. After the individual rating, the panel holds a teleconference. Here, applications with the lowest scores, usually a third, are culled using a "triage" approach. These are not dealt with in the subsequent panel meeting. However, if a reviewer objects and wishes to discuss a project previously cut off, it will be discussed in order to not disregard potential innovative projects.

The criteria to be scored at the Scientific Council for Medicine are the following: *Project*, which relates to the overall design of the project; *Feasibility*, including the materials, methods, resources, finances and time schedule of the project; and *Project Management*, including the skills and qualifications of the applicant and other people involved, as well as the research environment and collaborations. For com-

[1] From 2010-01-01, The Scientific Council for Medicine and Health

[2] Evaluation of the Panel Reorganisation in the Scientific Council for Medicine, Vetenskapsrådet 1:2010

petitive renewals, *Results* are also scored. The concern here is how far the project has, during the grant period, significantly contributed to an increase in knowledge in the field of medical research.

At the SRC, two different grading scales are used. There is a five-grade scale, VR grades, that is used by most councils and committees. In this scale, five is the best score and corresponds to outstanding, world-class research, while the figure one means insufficient quality. The Scientific Council of Medicine uses a 7-point grade; ranging from 7 for an outstanding, innovative project of highest scientific importance, to 1 for a project with no scientific relevance and/or of poor scientific quality. The current discussion at the SRC about grading scales, aiming to use a common scale for all SRC departments, relates to three different aspects of quality assessment. These are to compare quality within or between research areas and to show how results of research are beneficial for society, i.e. advocacy, when communicating the need for increased research funding.

## USA

The next speaker, Dr Andrea Kopstein, Director, Office of Planning, Analysis and Evaluation, Center for Scientific Review (CSR) at the US National Institutes of Health (NIH), began her presentation by agreeing there were many similarities but also some differences between Sweden and the US.

The NIH uses tax money to award grants, and peer panels are made up of external experts who are not obligated to participate and do so with only minimal compensation. In addition to it being an honour to review for NIH, other incentives include: delayed application dates for personal applications; and guaranteed review of an application within 120 days. NIH always looks to recruit the most highly qualified reviewers. The distribution of professors and senior researchers who review for NIH are from a variety of institutions and backgrounds.

The mission of the Center for Scientific Review (CSR) is to "see NIH grant applications receive fair, independent, expert, and timely reviews – free from inappropriate influences – so NIH can fund the most promising research". This, according to Kopstein, is achieved by peer review. Grant applicants do not know who their reviewers are. They can know who is on the peer review panel but the process is supposed to be as blind as possible. Approximately 75 per cent of all the research applications submitted to NIH are reviewed by the Center for Scientific Review.

The past decade has seen an increase in the number of research applications being submitted to NIH, from some 77,000 in 1998 to approximately 115,000 in 2009. This increase is largely due to applications submitted in response to initiatives associated with the American Recovery and Reinvestment Act (ARRA). ARRA is an economic stimulus package enacted by the US Congress in February 2009.

The peer review process within Center for Scientific Review (CSR) occurs in five scientific divisions, with 24-25 integrated review groups and approximately 240 scientific review officers (SROs). The first level of peer review takes place in a scientific study section (scientific review group). At the second level of review, the NIH funding Institutes and Centers use the overall impact scores assigned to the applications (and institute research priorities) to prioritise applications to fund. The applications are then reviewed by Institute Advisory Councils who make funding recommendations to the Director of those Institutes and Centers.

In recent years, the increasing breadth, complexity, and interdisciplinary nature of modern research necessitated a formal review of the NIH peer review system. This review resulted in a report with recommendations for enhancing the grant application peer review process. For the past year, NIH has been implementing enhancement recommendations such as a new scoring scale and shorter research grant applications.

With the new NIH scoring scale, applications are scored (1-9, with 1 as the best and 9 means poor) against five review criteria. The five core review criteria include significance/impact (does the project address an important problem or critical barrier to progress in field?); investigator(s) (are the programme directors, collaborators and other researchers well suited to the project?); innovation (does application challenge and seek to shift current research or clinical practice paradigms?); approach (is the overall strategy, methodology and analysis appropriate?); and environment (will the scientific environment

contribute to probability of success?). The new, 9-point grading scale also uses a new template-based technique, where the objective is to write evaluative statements and not overdo the grading.

The major goal of the change in the scoring process is that the rating scale is representative of scientific merit, and no other property of application. The review criteria include all aspects of the concept "scientific merit". Each reviewer's score is rated equally. Scoring is done in whole integers. Each application is assigned to three reviewers to score each criterion, write a short critique, and assign a preliminary overall impact score (priority score) for the application. In the US review process, applications are assessed on their overall impact, or the likelihood for the project to exert a sustained, powerful influence on the research field(s).

After the assigned reviewers post their preliminary scores, the applications are then brought for discussion to the review meeting or study section, where reviewers score all the applications in that particular scientific review group. Approximately the bottom 50 per cent of applications are not discussed at the scientific review group meeting. Applications discussed by any given peer review panel will also receive an additional overall impact score from all the reviewers in the scientific review group meeting. The CSR Scientific Review Officer then combines reviewers' critiques into a summary statement for each application. All summary statements and scores are passed on to the applicant and the potential funding NIH Institutes.

The previous NIH scoring scale – ranging from 1 (most outstanding) to 5 (lowest) – tended to cluster at the outstanding (fundable) range of the scale during review. The level of precision was too high, with a total of 401 possible values. In comparison, the new 1-9 scale means only 81 possible values. A pilot project using the new grading scale shows that the distribution of 720 applications during the summer of 2009 was balanced across the scale. The majority of the 40 reviewers involved were satisfied.

The Center for Scientific Review is also using alternative review formats, including electronic reviews (phone/web/video). This will improve the timing of the review process, as well as help keep costs down. As it stands, the majority of applications are reviewed in face to face meetings.

NIH is shortening page limits for competing applications to help reduce the administrative burden placed upon applicants, reviewers, and staff. This change seeks to focus applicants and reviewers on the essentials of the science that are needed for a fair and comprehensive review of the application. Starting in 2010, basic research applications may not exceed 12 pages in length. Other activity codes (for example Fellowship applications) have different page length requirements.

It has always been the intention of the NIH grants applications review process to rate, not rank, applications. Applications are to be rated independently of each other, not measured against others. When considering an application, the NIH looks at an applicant's track record as well as the innovativeness of the project. The Center for Scientific Review provides overall impact scores for all discussed applications but at the second round, the Institute looks at multiple things, including the applicant's track record.

Following her presentation, Kopstein commented on a question from Charli Eriksson, professor at the School of Health and Medical Sciences at Örebro University, about the importance of training related to the new grading scales. Is it possible for people to change their habits?

"This is a big shift for reviewers, especially the experienced ones," said Kopstein. "Both applicants and reviewers are used to 1-5 so it's a mental switch for both sides. It's too early to yet in the process to evaluate the impact of this change but we are assuming that training is essential for the success of this project."

Training can be talking to peer reviewer panels to introduce the new grading scale in detail and speaking to chairs of meetings to make them understand how to run meetings more appropriately.

## Norway

Next was Dr. Nina Hedlund, Special Adviser, Department for Marine Resources and the Environment at the Research Council of Norway (RCN), which receives around 5,000 research applications each year.

The review method at RCN is highly electronic – all applications have to be submitted electronically via the RCN website and all initial peer reviews take place using electronic systems. Like in Sweden and the US, experts work individually and in panels. Panels consist of three to six members and each panel looks at up to 20 applications. Each application is reviewed by at least two experts, and may be passed on to a third.

Following an individual rating, a two-day meeting is arranged for the panel to agree on its overall score. A written evaluation of each application is then passed on to the RCN, which bases its decision on this written statement as well as the expert panel's ranking of the projects.

Since the introduction of electronic application systems, there are about 20 different application types (research infrastructure, researcher project, user-driven innovation project etc.). Each has its own criteria, but use same scale system. What criteria to be used is selected centrally by the R&D group for each application type. Each activity involved in the project may decide which criteria will be evaluated by the panel.

Some of the criteria include: scientific merit, project management, research group, candidates for fellowships/grants (if relevant), feasibility, international cooperation, dissemination of results, relevance and benefit to society, environmental/ethical/gender equality/gender perspectives, relevance relative to the call for proposals, as well as budget limits (lower and higher). Not all criteria is evaluated by experts, which choose the most relevant for each project.

RCN uses a 1-7 scale, where 7 is exceptionally good and 1 is poor. The definition of the scale is the same for all criteria. However, RCN is now working on new, better definitions of the criteria and the scale. As it stands, some criteria are rated using an A-C scale and for others, a 1-7 scale will still be used. RCN is also trying to find specific definitions for the scale for each criteria.

The reason behind being more specific in the definitions for the scoring systems is that it is difficult to ensure that the scale is used in the same way by different panels.

Even more important than the use of the scale are the written words, as all evaluations are sent back to the applicants. When the evaluation process is completed, the names of experts involved are published within each knowledge area (not for individual publications).

RCN is currently investing in further development of its electronic systems.

## Finland

The final speaker in this session was Dr. Sara Illman, Science Adviser, Health Research Unit at the Academy of Finland, the country's main research funding organisation.

In Finland, evaluation of research applications is carried out by an almost exclusively international peer review panel. This is to avoid possible conflicts of interest in a small country but also to strengthen the international profile of Finnish research.

All applications that involve projects of more than 1 year will automatically go to peer review by an international panel. For shorter projects, funding decisions may be taken directly by the national research councils. Some 800-900 international experts are engaged every year.

The Academy of Finland receives around 4,200 applications per year, of which half go through peer review. In 2009, the Academy provided some €295 million in funding for high-level scientific research.

The panels arranged by the research council for health consist of 6-12 people and each member evaluates around 10 applications, or 30-45 applications per panels. Each application is reviewed by at least two members but sometimes also by an external reviewer.

Panelists are invited for a period of three years, if possible, which means that any given panel is likely to consist of 50 per cent members that have participated in earlier panels, and the other half are new. Before meeting their fellow panel members for an evaluation meeting, each panel member writes a draft statement and gives a draft score to the applications that have been allocated to him/her. There is no triage process, all applications will be discussed. Based on panel discussions, a final consensus score

is given on behalf of the entire panel. The consensus statement is based on the draft statements, and the written feedback is sent back to the applicant after the evaluation.

The final decision is made by the Academy's research councils, which base their decisions on statements from all the panels, taking into account science policy issues and budget limits.

Until 2007, the scale of rating used to be 1-5, with 5 being awarded to outstanding proposals and 1 being given to poor proposals. Panels were encouraged to use the entire scale but, in practice, the panels sometimes used an additional system of + and - to separate applications, especially those given a 4 rating. As a result, a new scale was introduced in 2008. The new scale ranged from 1-6. This change in rating criteria resulted in a better balance between the higher-end ratings.

The Research Councils of the Academy are not bound by the grades given by the panels but will obviously respect this rating highly. In general, all 6-rated applications are approved unless there is a major problem with the project or application. Depending on the funding form, many 5-rated applications will be funded, and usually also some of the applications given a rating of 4. For a project rated as 3 to be funded there would have had to be some misunderstanding by the panel. Lower-rated applications may be funded if there are strong reasons or real benefits to Finland.

Stefan Björklund, Professor at the University of Umeå, asked whether having only two members read each application poses a problem. Can there be consensus; isn't it easier to rank an application when everyone has read it?

"The panels don't rate, they only rank applications," Illman advised, adding that reviewers must write about the application in their own time. "The Academy picks the best reviewers for each application, and strong and weak parts are identified and discussed by the panel."

# SESSION 2: THE SCIENCE OF GRADING SCALES AND THE EXPERTS' OPINION

The first expert presentation in this session – aimed to present recommendations for the future grading scales – was Professor em. Marie Åsberg from the Department of Clinical Neuroscience at Karolinska Institutet and StressRehab KIDS at Danderyd's Hospital.

Having worked as a psychiatrist since 1962, Åsberg is very familiar with making psychiatric diagnoses with the help of rating scales. In her presentation, she explained that the first rating scales were constructed by meteorologists. Measurement using rating scales have been problematic as everyone has their own set of experiences that affect the rating.

What exactly is a rating scale and how complex must it be in order to be called a rating scale? asked Åsberg. This question has been debated since the 1940s.

"Measurement is the assignment of numerals to objects or events according to rules," Stanley Smith Stevens established in 1951. A later definition by Nunnally & Bernstein (1994) states that measurement "consists of rules for assigning symbols to objects so as to (1) represent quantities of attributes numerically (scaling) or (2) define whether objects fall in the same or different categories with respect to a given attribute (classification)".

Rating scales are often used within psychiatry. For example, the Likert scale is prevalent. It is usually a summated scale, where different items are scored on ordinal scales. The scales usually contains definitions of individual items (=symptoms) and cues for the different scores.

"It is quite easy to design a scale, but quite difficult to design a good scale," Åsberg commented.

When constructing definitions and cues in rating scales it is essential to consider the following:

- Clarity: definitions should be simple, unambiguous, short
- Relevance: must be of cue to item
- Precision: cue must exist in correct rank on the rating scale
- Variety: cues must not contain the same wording
- Objectivity: ethical or social evaluation must be avoided
- Uniqueness: avoid general terms e.g. good, fair, poor

There are also a number of classic rating errors, explained Åsberg. According to J P Guilford (1954) these are:

- The Error of Leniency – risking bias due to existing relationship or knowledge (colleagues etc.)
- The Error of Central Tendency – leaning towards the middle of the rating scale
- The Halo Effect – when the expert has a general impression of something and lets this affect the rating
- A Logical Error – experts making unconscious errors due to their own logic, that two things belong together (such as physics/mathematics)
- A Contrast Error – such errors are based on the expert's own characteristics, based on themselves. If the expert is aware of this, they can avoid making the mistake, which is an error in itself
- A Proximity Error – when items are closely related or close to each other, they tend to colour each other and this can open for errors

Åsberg has been involved in constructing two different rating scales: Comprehensive Psychopathological Rating Scale (CPRS), which was initiated by the then Swedish Research Council in 1978, and Montgomery-Asberg Depression Rating Scale (MADRS), which is often used for depression, both in Sweden and internationally. In fact, it is the second most used scale internationally.

Before evaluating the success of a rating scale it is essential to check how two raters or a group agree with each other, and what the inter-rater reliability is. If raters are motivated, experienced and well-trained, this often works very well. Consequently, a good scale permits very high inter-rater reliability but does not ensure it, unless raters are motivated and well-trained.

Evaluations of the so-called GAF rating showed that it had a high accuracy – in the range of 0.77 to 0.86, compared to an average reliability of around 0.60. It showed that experienced raters do better, i.e. have a higher level of reliability. Another finding was that reliabilities were clearly lower when rating clinical impact than when rating scientific quality.

Validity is much more difficult as you have to look at how an individual rates him or herself compared to the rater. There is often discrepancies between the rates given by an individual, compared to when he or she is part of the group.

"Thus, rating is not only a measurement but also a social experiment," Åsberg explained. "All sorts of things happen within a group, and situations are different for each group. You can never predict what will happen or what dynamics will be prevalent within a given group."

Arne Jarrick, Secretary General for the Humanities and Social Sciences, Swedish Research Council, asked about central tendency errors in reliability tests. Do they exist these days?, he wondered, adding another question on the performance of experienced raters vs. consensus errors. Do even rating numbers help in this aspect?

Åsberg said that central tendency errors have been replaced by error of lenience. This is usually due to colleagues feeling uncomfortable when using low ratings on their peers. She added once again that the rating process is an important social experiment, and that each rating situation is different.

Karin Forsberg Nilsson asked whether we want trained reviewers that perform better? What benefits and/or risks exist with a panel having consensus from the start?

Håkan Billig commented that a few years ago, there was an issue with different ratings between the Scientific Council of Medicine at the Swedish Research Council and a foreign research council. But even if the ratings differed, the correlation between individual ratings and group ratings had a high correlation, which was "very pleasing" according to Billig.

The next presentation focused on the psychology of grading scales. Professor Henry Montgomery, from the Department of Psychology at Stockholm University specialises in cognitive psychology.

"Grading scales are natural," Montgomery commenced, as grading is built into any language. We rate things and experiences as excellent, very good, good, OK, and so on. In the academic world as well as in everyday life, institutionalised grading has been used for a long time (rewards in relation to merit, grading of social groups, school grades, grading in sports and music competitions etc.). There is often a high level of inter-judge and international agreement related to certain grading (e.g. facial beauty, where $r > 0.90$).

Peer reviews often produce low levels of agreement (correlations of 0.19 - 0.54 for manuscript review and 0.18 - 0.37 for grant reviews). However, "mature" sciences often have a higher agreement rate (indicated by high acceptance rates of manuscripts submitted to high quality journals, e.g. 91 per cent in astrophysics).

An 1956 article by George Miller concluded that the optimal number of response categories is seven, plus minus two. Seven is the perfect amount, that people can distinguish between and keep apart. If needed, the number seven can be increased by supplementing response categories with + or -.

One must also decide on the scale level, whether it is ordinal, ratio, or absolute. Verbal labels may be included but the absolute level must be the same across all areas so that it can be compared.

The so-called Category-Ratio (CR) scale, constructed by Gunnar Borg combined level and labelling considerations. The CR scale has been considered the 'ultimate rating scale' as it combines everything.

When using multiple formats, consistent response formats are essential. The afore-mentioned 'halo effect', when the expert has a general impression of something and lets this affect the rating, is very strong. Preconceived ideas have an overall value when influencing gradings of specific aspects. Groups often seek coherence so there may be discrepancies between individual and group ratings.

Response compatibility must also be taken into account – it is easier to assess quantitative than qualitative aspects with quantitative scales. However, this poses a risk that quantitative aspects (e.g. number of publications) are overweighed when quantitative scales are used.

What reference points do raters have, and are they implicit or explicit? Many reference points are possible when grading research applications and not all of them are explicitly stated. Examples of reference points include previous ratings of research proposals from the applicant, top quality proposals within the country, top quality proposals on an international level. Depending on the choice of reference point the same proposal can be graded very differently.

Montgomery showed an example: how financially content citizens in certain countries in Europe are. Latvia, for example, proved to have a higher level of contentment than the Italians. Latvia had recently experienced a financial boom but were still poorer than Italians, but this did not affect their level of contentment.

Also, the frequency of applications judged to have lower and higher quality of an application up for rating is important. The higher the frequency of inferior and the lower the number of superior applications (i.e. the higher the ordinal rank is), the more positive will the grading be.

The conclusion is that when you rate something, a fair grading should reflect a reasonable compromise between distance to reference points and the frequency of applications judged to have lower and higher quality than the proposal to be judged.

It must also be established what the evaluator wants to achieve with the rating? Is the evaluator reporting an evaluation of a research proposal and nothing else? Or does he or she also want to achieve something with his or her rating besides just reporting an evaluation, such as helping, encouraging, demonstrating his or her expertise or fairness, express his or her values or (in the worst case scenario) to stop competitors. In any case; to evaluate is to act since there are consequences (legitimate or illegitimate) involved with different levels of evaluations.

A current rating scale applied by the Swedish Research Council's Scientific Council (used in the evaluation group for psychology) ranges from 1 to 7, where 7 is a very high measuring and 1 is low. These ratings are given to applications on five criteria; the general importance of the research team, the scientific importance of research questions, the originality and potential for innovating research in the application, the theoretical relevance and anchoring in the application, and the method's and data's adequacy for the research project and the availability of data.

There is also an additional 1 to 5 rating scale, measuring the overall scientific quality of the application. "This is very confusing," Montgomery commented, "and as a member of the evaluation group it took me a long time to understand the different formats of the specific scales and find consensus."

Montgomery warned there is also a high risk of halo errors as a group seeks coherence. He advised on various possibilities of improving scales, including a consistent response format, and more concrete scales. Another suggestion is to introduce independent evaluations of specific aspects, in order to counteract coherence seeking. "I am surprised by the lack of research into this," he said. "International cooperation on this subject is needed."

Stefan Björklund, Professor at the University of Umeå, commented that at Umeå's medical programme, experts only use a pass or fail grade to rate projects. "A 7-grade scale creates lots of extra work," he said, "so can a simple pass or fail system be something for the future. Is it good enough to rate projects?"

Montgomery wondered how the size of funding is decided, with a simple pass or fail system? Using rating scales, the projects with the highest ratings get more money.

The final speaker in this session was Professor em. Elisabeth Svensson, from the Swedish Business School/Statistics at Örebro University. She presented statistical aspects on quality assessments in peer review.

Svensson is one of very few statisticians with rating scale statistics as a research field. She has developed statistical methods for evaluation of data from scale assessments. These are applicable to evaluation of quality of ratings and of scales, but also to evaluation of change.

The use of rating scales in peer review of research generates ordered categorical data, also called ordinal data, Svensson explained. This type of data indicates an ordering only, meaning that any relabelling of ratings must not affect the result of analysis. S.S. Stevens defined these properties that only monotonic transformations are permissible in data from rating scales. Consequently, Stevens stated that the "arithmetic mean is not a proper statistic for an ordinal scale, although it is often used in averaging such ordinal values as scores on tests and grades in courses". This means that rank-based statistical methods, such as calculating median score, quartiles and other centiles are appropriate for description.

Svensson demonstrated that different profiles of multiple scoring of research quality – for example by the variables research purpose, competence, methods – can give identical sum scores. She presented various alternative methods for defining a global score of quality of research, such as hierarchical conditional score, median score, or some other rules defined in advance.

The ultimate goal in assessment of research quality is to have high quality in the peer review grading system. What is the validity of the grading system, and are the reviewers reliable?

Svensson demonstrated her statistical methods that are developed especially for ordinal data from rating scales by means of data sets representing inter-scale comparison and inter-rater reliability, respectively.

Inter-scale comparisons are used in validity studies. She demonstrated that the use of a visual analogue scale (VAS) is least reliable for research. This is because the 101 possible positions on the grading line are not interpretable in words and are not comparable with assessments on a verbal descriptive scale for the same variable. Svensson's research has also shown that condensing data from VAS assessments to a discrete number of categories, say 10, or from a 7-point scale to a 5-point scale, for example, will result in data that are biased the origin.

Evaluation of inter-rater reliability deals with agreement in grading the same application between reviewers. If the reviewers do not agree completely, why do they disagree?

By Svensson's method it is possible to identify and measure the components of systematic disagreement separately from an individual based disagreement, when present. This means that it is possible to identify reviewers that systematically use higher scores than others; a difference in scores between male and female reviewers, between experienced and less experienced reviewers, etc. will be revealed if present. Poor quality of scales and of the operational definition of the variable of interest will produce a high level of individual based disagreement.

Svensson stated that it takes time to implement new statistical methods, and to become accepted among users of statistics. It is easier to stick to traditional methods, whether appropriate or not. The use of correlation coefficient in reliability studies, the use of sum scores, and the use of parametric statistical methods of analysis are common misuse of statistical methods for data from rating scales. It is unethical when decisions are based on such misuse, Svensson concluded.

# SESSION 3: PANEL DISCUSSION

The subsequent panel discussion featured all the previous speakers, and was moderated by Karin Forsberg Nilsson. The topic for the discussion was "How can funding organisations optimise their scoring in quality assessment?"

Karin Forsberg Nilsson opened: "This has been an excellent opportunity to listen to presentations and discussions among peers. It is a difficult topic but we are doing our best. We are supposed to fund research of the highest quality, that is in the nature of what we do. Therefore the ratings will also need to be of the highest quality."

The panel then summed up the lessons learnt during the seminar and how these findings can be used in the various organisations.

The Research Council of Norway will continue using its 7-grade scale, stressing that the verbal description of each criteria is important both for reviewers and applicants. The NIH will also keep its current 9-grade scale. "The use of verbal descriptors is interesting, and it will be interesting to see how these are used," Andrea Kopstein commented.

The Academy of Finland will also stay with its 1-6 rating scale. Sara Illman added that based on her experiences from the seminar, the Academy would work on introducing short and simple verbal formulations.

But how to measure innovation?, the panel asked. Should criteria be separated and should separate scales for separate panels be used?

Someone asked whether the track record of a scientist and what they have published should be used in the application process?

"We need different quota in the application process. More junior researchers don't have scores in the same way as their senior peers," said Andrea Kopstein. "At NIH, we give junior researchers scores in meetings etc., in order to build up their profile. Panels are also told to fund some of the juniors, giving them special attention. Thus, there is no need for a separate budget."

Håkan Billig added that within medicine fundings at the Swedish Research Council, 40 per cent is earmarked for people 10 years or less into their doctor's degree. "This makes them able to play on the same level as more experienced researchers."

In Finland, there are no separate rules for junior researchers but these are always pointed out to the reviewing panel "and should be encouraged", according to Sara Illman. "Innovation is included in our criteria, and great true potential is pointed out."

Charli Eriksson suggested: "'What is good science?' That is how we should measure projects. Metric methods could be used instead of panels, but innovative projects should be funded."

NIH uses an overall impact priority score to identify the most innovative ideas. "We try to inspire reviewers to find innovative projects," said Andrea Kopstein.

The issue of comparing old data and extending the use of validation kept within the scientific community was brought up. "Why haven't we compared data from the 1990s to now?" Håkan Billig asked. Many of the panel members agreed that retrospective studies would be useful. Being able to look at both applications that were funded and those that weren't could give interesting insights into the success of grading scales.

The Research Council of Norway has little information about projects that were not funded. The NIH carries out reporting every year, but also doesn't consider applications that weren't funded. Continued assessment of peer reviews happens at NIH every year, said Andrea Kopstein, adding "we have a constant outlook to improve".

Arne Jarrick suggested picking a sample of research projects that have just received funding – some of which would have been considered 'excellent' and some who just passed – could be one evaluation method. "Five years later, the research council could go back and measure them again to see if the same rating still applies, asking whether the outcome of the research matched the ratings and assessment?"

Elisabeth Svensson agreed that it is a "very good idea to see outcome of applications and compare after three or five years. Measure the substance of applications is very interesting."

Håkan Billig thought this may be a good idea but wanted a different time scale for measuring a project's outcome. However, Jarrick warned about the risks involved in waiting too long, as the values and rating scales of the research community may have changed.

The Swedish Research Council holds discussions on how to improve processes, how to avoid conflicts of interest etc. every year. "Comparing applications has been going on ever since peer reviews have existed," commented Håkan Billig.

Everyone agreed that it is the duty of any research-funding organisation to continuously look at their systems in order to follow best practice. "As long as there isn't an ideal way to process applications we must continue to assess ourselves," said Sara Illman.

On the topic of what checks are carried out to evaluate if the best science was picked, Karin Forsberg Nilsson asked what follow-up processes of old data exist?

The panel members were in agreement that comparing a project's progress with the wording and goals in the project application was not a good idea. For example, experienced panel members have a huge benefit when writing applications and are usually very successful when applying.

A research-funding organisation must only ask project leaders if the project is going according to plan, according to the timescale, goals etc. stated in the reporting forms. The research council must have faith in the response. "However, any changes to the project must be reported immediately, as new negotiations and possible new fundings may be required," added Nina Hedlund.

Håkan Billig agreed: "A comparison between project progress and application goals must not be done. In basic research, it is detrimental to impose rules about following a project plan. Of course, certain rules must be applied but I would be worried if 50 per cent of our project leaders were following their project plan!"

"Research develops as it goes along and we can't hold applicants to an exact route. Researchers must not feel that this is another layer of bureaucracy," Karin Forsberg Nilsson added, before asking for final comments on the seminar.

Lil Träskman Bendz summed up the seminar: "It is very useful to sit down and listen to the various experiences between countries. It is important to have similar ratings and evaluations of research, in the Western world at least. Research should be followed up as experienced members tend to have similar ways of rating. There are many different reasons for this – the halo effect, they know each other already etc. As you can see, there are several problems with validating ratings."

Mats Ulfendahl was worried about the discussion topic: "As a panelist and researcher, I'm scared that these numbers are used for other things. What are the scales supposed to be used for? To calibrate different areas? Is it possible to use the same scale for everything?"

Karin Forsberg Nilsson summed up by saying that the last question was a good outline for a follow-up seminar and thanked all members of the panel and audience for attending the seminar.

# SAMMANFATTNING

Diskussionen kring kvalitet och skattning av kvalitetsparametrar i forskning är aktuell i många sammanhang. Det handlar för det första om bedömning av ansökningar, d.v.s. skattningsskalor som instrument att bedöma forskningens kvalitet. Den andra aspekten är om och hur skattningsskalor kan användas för att jämföra kvalitet inom eller mellan forskningsområden. Det finns ett värde i sig för forskningsfinansiärer att jämföra områdens kvalitet som en del i arbetet med att stödja den bästa forskningen och för att kunna göra s.k. horisontella prioriteringar. Det tredje skälet att använda välutvecklade bedömningsskalor är kommunikationen med omvärlden om forskning och forskningsfinansiering.

Denna skrift är ett referat av ett seminarium som diskuterade användningen av betygsskalor vid peer review-bedömning av forskningsansökningar. Seminariet organiserades av Vetenskapsrådets ämnesråd för medicin och hälsa. Syftet var att bidra till den fortlöpande diskussionen om kvalitetssäkring av Vetenskapsrådets bedömningsprocesser.

Seminariedeltagarna ansvarar själva för de påståenden och slutsatser som förs fram i de olika bidragen. Bidragen representerar således inte nödvändigtvis Vetenskapsrådets ställningstaganden.