Assimilation and Contrast Effects in Sequential Judgments

Dissertation im Fach Psychologie zur Promotion
in der Fakultät für Verhaltens- und Empirische Kulturwissenschaften
der Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Dipl.-Psych. Thomas Haar

2004

Dekan:	Prof. Dr. Klaus Roth
Berater / 1. Gutachter:	PD Dr. Henning Plessner
2. Gutachter:	Prof. Dr. Thomas Mussweiler (Köln)

Dedicated to Claudia & Coralie

The two nicest little women that have ever breathed

-

(I now declare this bazaar opened!)

Acknowledgements

I would like to thank:

Ali der Tanzmaus, Alina, Amy, Andreas G., Andreas St., Andrei, Anja D., Anja Sch., Anna L., Anna T., Antoine, Arie, Armin, Arthur, Bäcker-Katrin, Barbara, Basti, Beate, Bernd, Birgit, Bob F., Bob the Hero, Calvin, Cello-Andreas, Cello-Jonathan, Chris L., Chris Sch., Christian R., Christiane, Christina, Christine, Christoph M., Christoph R., Christoph S., Claudi, Clemens, Connie, Coralie, Cornelia, Cordel, Csaba, Dara, David, dem Eiermann, dem King of Queens, dem Iltis, dem Kaiser, dem Meister, dem Nobody, dem Schleicher, der Weberwurst, den Lehmanns, den Mensafrauen, DFG, Dino, Dirk, Ditsche, DLH, Doris, Dr. Christian B., Dr. Christian G., Dr. Christian U., Dr. Günther, Dr. Henning, Dr. Mavridis, Dr. Simon, Dunga, EAESP, Edo, Emilie, Emma, ESCON, Eva, Family Kanie, Flo, Franka, Frau Dr. Branke, Frau Schmitt, Friederike, Geoffrey, Gerd, Gerit, Gideon, Götz, Gregor “Kuhgott" W., Janne, Hanna, Harriet, Hasan, Heike, Heinz, Helge, Helen, Hendrik, Henni, Henning, Herrn Bothe, Herrn Kaiser, Herrn Schollmeyer, Hobbes, Horst, HSC, HSV, Ines D., Ines P., Ingo Dimingo, Iris, Isabell, Jan-Oliver, Jasper, Jens, Jens-Erik, Jessika, Joanette, Joe, Johann von, Johanna, John, Jonathan, Julian, Juliane B., Juliane D., Jürgen, Kai, Kaiser, Käpt’n Blaubär, Kati, Katja H., Katja R., Katrin mit Kind, Katrin ohne Kind, Katrin J., Karin, Ken, Ketten-Jan, Kirsten, Klaus F., Klaus Sch., Klinsi, Konne, Kröning, Linda, Line B., Line J., Lili, Majid, Malte, Marcel, Mareile, Maren, Marion, Manni, Mario, Markus, Martin, Matt, Matthias, Matze, Meyer, Michule, Miriam Sch., Miriam Sp., Mom and Dad Hanley, Mr. Aufziehvogel, Mutter, Murat, Nana, Nici, Nina, Norbert, Oliver, Olivia, Olivier, Oma, Onkel Andi, Onkel David, Orhan, Özlem, P&S, Pascal, Paulo, Peppi, Peter F., Peter K., Pfr. Mohr, Pfr. Tonka, Phil, Philipp, Preussi, Prof. Dr. Abdul Nachtigaller, Ralf E., Ralf R., Rasi, Rebecca, Rocko, Rolf, Rohel, Rudi, Rüdiger, Rumo, Ruxi, Sabine C., Sabine H., Sarah, Saskia, Schildkröte, Schulze, Sebastian, Seppl, Silke, Simon, Sören, Sprolli, Stanna, Stas, Stefan H., Stefan J., Stefan K., Steffi, Steino, Stirl, Susanne, Tante Erika, Tante Linda, Thomas B., Thomas M., Tilmann, Timon, Tina, Tini, Tobi K., Tobi M., Tobias, Tom, Tommie, Ulli, Ulrike, Uschi, Vater, Volker, Walter, Willy, Wolfram, Yaël, Yvonne.

Contents

Summary

Based on my experiences both as a student actively taking exams and as an assessor passively journalizing exams, I was interested in the influence of prior on subsequent evaluative judgments in a sequential judgment situation. In most (written or oral) exam situations, the performances of different students are judged in a sequence. The basic idea, I started out from, was that in this case performances of different students to be judged during an exam are compared with each other; more precisely, I expected the performance judgment of a target student to be influenced by the performance judgment of the prior student. Thus, the prior student was expected to function as a comparison standard for the target student to be judged in an exam situation. This might have an assimilative as well as a contrastive effect on the judgment of the target student. In a first line of studies it was tested *whether* and *in what direction* judgments of prior performances would influence subsequent ones. Participants were in the role of a teacher and had to grade the performance of two students during an exam based either on written or oral protocols. The performance of the first student was manipulated to be either good or bad; additionally, the focus of participants was manipulated to be either on similarities or on dissimilarities via an ostensibly unrelated priming task. The results show that prior judgments may have an assimilative as well as a contrastive influence on subsequent judgments, dependent on the comparison focus of the judge. As suggested by the *Selective-Accessibility Model* (SAM, Mussweiler, 2003a), a similarity focus made assimilation effects more likely whereas a dissimilarity focus made contrast effects more likely. Over all three studies, assimilative influences were stronger than contrastive influences. This is in line with the SAM that describes assimilation effects as the default influence of a standard in comparative judgments. In a second line of studies, assumptions derived from the SAM concerning the *processes* underlying comparison effects on judgments were directly tested. The model suggests that prior to a judgment the target to be judged is compared with a given standard. During this comparison, judges follow and test either a *similarity* or *dissimilarity hypotheses* – employing a certain (positive) test-strategy – and will, thus, selectively activate information on which the final judgment of the target will be based. Participants were again in the role of a teacher and had actively to test two (or three) students in a

virtual exam, using a computer-based simulation. Participants had to ask questions and received answers from the virtual students they had to grade in the end. The same variables as in the first line of studies were manipulated. The results show that participants used a positive test-strategy to test their hypothesis regarding the target student. They asked more difficult questions when expecting a well performing student than when expecting a badly performing student. The influence of this test-strategy was more pronounced at the beginning compared to the end of an exam, suggesting that participants were able to integrate (disconfirming) feedback and to adjust their hypothesis accordingly. To conclude, the first part of my work shows that comparisons may influence evaluative performance judgments in a sequential judgment situation. The second part gives first direct evidence for the assumption (derived from the SAM) that these effects are caused by an underlying process of (positive) hypothesis testing.

Chapter 1

-

Exams put to test:
Assimilation and contrast effects in performance judgments

1.1 Introduction

Once upon a time, there were three musketeers trying to graduate. In the year 2001, two friends of mine and I started our odyssey through our final exams. We had to face seven oral exams until the marathon would be over in the summer of that year. Fortunately, we spent a lot of time together, preparing, learning, and motivating each other in times when almost all hope seemed to be lost. Moreover, we managed to time our exams so that we were always tested on the same day, in order, one after the other. So, needless to say, we all joined the first candidate who had to walk the plank. When the teachers – surprised to see three of us waiting – asked who would be the first to take the exam, we soon found out that it was up to us to decide the order of our three exams. After the first time this happened, we decided that it would be wise to spend some time on deciding which order would be best for all of us. It seemed obvious that we would be compared with each other and that the performance judgment of the first exam would influence the following performance judgments.

Yet, it was not so clear, in what direction this influence would be. We knew from our learning sessions that each of us was an expert in some field; so, we could rank order our expected performances for every single exam. Therefore, the question was, whom we should send in first. If we sent in the best one of us, would his shining performance blind the teachers and make the other two look better? But we also thought about the danger that this strategy could backfire. What if the two following candidates were looking worse compared to the excellent performance of the first one? Should we, therefore, send in the worst one of us first? In that way, he would not look bad compared to the others and the two others could even benefit from the contrastive comparison with his performance. We discussed this topic before every single exam, but we could never decide and, therefore, tried both strategies with different results. Anyway, we graduated and all these thoughts and discussion about sequences in exams were soon forgotten.

However, when I decided to continue in academia as a scientific assistant, the very same questions soon came back to my attention. One of the tasks I had to fulfill

as a scientific assistant was to write protocols as an assessor in oral exams. This time I could see the whole exam situation from a different perspective, which offered me some new insights. The most important thing I noticed was that teachers did, indeed, compare the performances of different students with each other. So, the question popped up again in my mind in what way prior performance judgments would influence the judgments of subsequent performances. And how could these effects be explained, not to forget how it was possible to push them in the wanted direction? Therefore, I started with my dissertation project – reported here – in order to find an answer to these questions haunting me.

1.2 What it is all about

As described in the introduction, I started my work with the question if the judgment of prior performances does systematically influence the judgment of subsequent performances in oral exams. Judging the performance of students in oral exams is a special judgment situation for some reasons that I would like to outline in the first section, followed by a description of the core questions guiding my dissertation project.

First of all, judging the performance of students in oral exams is – in most of the cases – an example of a sequential judgment situation. These are situations where a judge has to judge at least two different targets in a sequence on the same dimension. Another example for sequential judgments may be the judgment of candidates applying for a job. Although sequential judgments may include any kind of judgment dimension, I will focus on *evaluative judgments* that are relevant for the exam situation. In this situation, the performance of the students is judged on an evaluative dimension (from excellent to insufficient).

Secondly, judging the performance of students is a complex, relevant and important judgment situation. Compared to this, much of the research in the field of social psychology on judgments focuses on rather simple, irrelevant or unimportant judgment situations. Taking an exam is a situation relevant to most people, because it is a situation that almost every one has to face at least once during his educational or academic life. The performance judgment or grade is also very important for people's future academic or job prospects, especially in times as today where the job market is relatively tight for new applicants. It is important for me to note that with my work I want to focus on a more complex judgment tasks, an important feature of most applied judgment situations.

Last but not least, because the performance judgments during an exam have important consequences for the people being judged, it seems natural that these judgments should only be based on the quality (and quantity) of the performance itself. Additionally, the exam situation should be the same for all students being tested. Following this call for objectivity, all students should be judged facing the same situation, and the judgment should be based on the same judgmental rules. Only if this is accomplished, it is possible to assign the performance judgments (i.e., grades) an absolute meaning and to compare different students' performance judgments of – coming from different schools and age groups, judged by different teachers – with each other. It is important to note that, therefore, the performance

judgment of a student should not be systematically influenced by prior judgments. Nevertheless, if there were such influences, they are not wanted and will, therefore, be seen as biasing the judgments and will be termed judgmental biases.

The first question I try to answer with my work is whether there are systematic influences of prior judgments on subsequent ones in a sequential judgment situation. My basic idea is that the performance judgment of a target is compared with the performance judgments of prior targets and that this may influence (or bias) the judgments in an unwanted way[1]. As outlined in the introduction, this influence may possibly be in two directions: Subsequent judgments may be biased towards (assimilative influence) or away (contrastive influence) from prior judgments. So, I do not only want to find out if there are systematic influences of prior performance judgments on subsequent ones, I am also interested in the direction of these influences, assuming that there are any.

The second question is concerned with how these influences can be explained, again, assuming that there are any. I already mentioned the idea that the 'process of comparing' prior performance judgments with subsequent ones may play an important role. This is at least what I observed as an assessor in exams. So, the second goal of my work is to take a close look at the underlying processes, which may explain or mediate the assumed influences of prior judgments on subsequent ones.

[1] Although my work only focuses on the influence of prior *judgments* on subsequent ones, it may easily be extended to the influence of the mere perception of prior performances on subsequent performance judgments or, again, on the mere perception of subsequent performances.

1.3 Sequential performance judgments in exams

A first approach to answer the questions just outlined is to search through the existing body of literature. Therefore, in the next section I will present an answer to these question based on existing findings and theories. First of all, I will discuss the question whether it makes sense to expect judgmental biases (i.e., judgments being influenced by unwanted variables) when making evaluative performance judgments during an exam. Second, I will give an overview of potential biasing variables that may be important for the special situation of sequential judgments. Here, research will be presented that shows judgmental biases for the case of sequential judgments. Third, I will focus on the literature regarding 'social comparisons' since I expect comparisons to be an important driving force behind biasing effects in sequential judgments. Here, I will especially present the *Selective Accessibility Model* (Mussweiler, 2003a), which will be used as a guiding model for the empirical part of my work.

1.3.1 Subjective biases in objective judgments

Everyone has to take exams during their life, either as a pupil at high school, as an apprentice during his apprenticeship or as a student at university. For most people taking an exam is a rather unpleasant experience, yet, the resulting performance and ability ratings have an important impact on our future careers and life. Therefore, it seems important to ensure that these ratings are as correct as possible, thus, ensuring that comparisons between different performances are valid or fair. In the following section it will be discussed whether this obviously necessary demand for validity or objectivity is satisfied by judgments in exams.

The axioms of the *classical test theory* (Gulliksen, 1950) say that every observed value of a person measured or tested can be decomposed into the *true value* of that person and an additional *error term*[2]. The goal of any test should be to measure as much of the true value of a person and as little error as possible. Examinations can be seen as *tests* trying to measure abilities of pupils or students. Examiners should, therefore, try to make sure that the size of the error component in exams is as low as possible, especially, since these ability judgments are an

[2] Following the connotation used above, this error term can be seen as equivalent to a judgmental *bias*.

important factor for people's future careers. To ensure that a test provides the true values of the persons tested, it has to be objective, reliable, and valid.

Over the years the objectivity, reliability, and validity of exams has been the focus in a large body of research, coming up with a number of factors feeding the error terms of the observed values (see Wherry & Bartlett, 1982). Since examiners have to rate the written or oral performance of students, much of the research is concerned with the objectivity of the examiner or of the exam, respectively. Many examiners seem to have a stable style of grading or rating, such as a lenient or stringent rating style, or the tendency to prefer or to avoid extreme ratings. For the performance of each student is usually only rated by one examiner, these rating styles add to the error term of the performance ratings. For example, Raymond, Webb, and Houston (1991) investigated oral examinations administered in a medical specialty, examining over 350 test candidates examined by 60 raters over a 3-year period. They found 25 % of the ratings to be significantly biased by lenient or stringent rating styles. In their work they also developed a method to correct this bias.

Rating exams can also be seen as *social judgments*. The *social cognitive approach* (starting for example with the work of Solomon Asch, 1946) as well as the more general *social psychological approach* to person perception (starting for example with the work of Gordon Allport, 1937) both try to understand how social judgments are formed and – more importantly for my work – how accurate social judgments are achieved and what variables might influence judgmental accuracy. In his *Realistic Accuracy Model* Funder (2003) describes the stages necessary for a judge to come up with an accurate judgment. For a start, the person to be judged has to perform behaviors relevant or diagnostic for the dimension to be judged. Second, the relevant behavior has to be available to the judge. In the case of an exam one can assume that these two necessities are given. The exam situation should allow a student being tested to show at least some of his abilities (e.g., by asking questions to test the knowledge of a student) and the performance of this student is also available to the teacher because "the availability step simply means that the judge must be present" (Funder, 2003, p. 122). The following two stages seem more important when it comes to accurate judgments during an exam situation. In the third stage the judge must successfully detect the available performance behavior and must finally – in the fourth stage – use this information to form a judgment. This last stage includes the interpretation and integration of all the information available in the situation including the information that the judge may retrieve from memory.

Obviously, these last two stages offer quite a fertile ground for biases to breed and grow. Indeed, social cognition research has been demonstrating many different

examples for judgmental biases in recent years. Funder (2003) comes up with a number of 39 different biases or errors reported in the social cognition literature (e.g., confirmation bias, halo effects, fundamental attribution error) and he also makes some critical remarks:

> The errors form a long list but no particular pattern. Typically, they are studied one at a time; some are effectively the property of particular labs or investigators. Singly or together, they do not yield a general theory of judgment, nor do they explain how accurate judgment is ever achieved. (p. 118)

With this critique he calls for theoretical frameworks that try to explain these biases rather than finding more and more singular empirical evidence of judgmental errors. Such a framework is, for example, the cognitive-ecological sampling approach to judgment biases (Fiedler, 2000). The approach assumes that many judgments are based on samples that are either drawn from memory or based on searching the environment. These samples are "virtually never random" (Fiedler, 2000, p. 660) and may, therefore, be biased in many different ways. For people lack the awareness and the ability to correct these biased samples, they base their judgments on the sampled information as if it was drawn randomly. With his approach Fielder gives an explanation for many well-known judgmental biases such as the illusory-correlation bias, the confirmation bias, and the phenomenon of base-rate neglect.

Still, the sheer number of judgmental biases that exists in the literature also tells us that there are many barriers on the way to an accurate judgment. The cause of these biases may lie within the person judging – as described above –, but "in many cases the variation of judgments is due to a change of judgmental context" (Bless, Schwarz, & Wänke, 2003, p.180). There are many factors that may possibly influence the judgment of a teacher grading exams. Research of judgmental biases in other areas suggests that the mood of the teacher (e.g. Schwarz & Clore, 1983)[3] – the way the students to be judged are dressed (e.g. Forsythe, Drake, & Cox, 1985), whether they are wearing glasses (e.g. Edwards, 1987), whether they are sitting in an upright position (e.g. N. A. Murphy, Hall, & LeBeau, 2001), whether they are smiling (e.g. Reis et al., 1990), or their physical appearance in general (e.g. Dipboye, Fromkin, & Wiback, 1975; Eagly, Ashmore, Makhijani, & Longo, 1991) may also influence performance judgments during an exam.

[3] Which may again be influenced by factors like the weather or the temperature of the room.

Of course one may ask whether these context factors may have a biasing effect even when making such important judgments as grading (final) exams. When judgments have important consequences, judges should be highly motivated to be accurate. In addition, teachers making the judgments are experts[4] in grading exams. It seems, therefore, unlikely or at least questionable to expect judgmental biases even under these circumstances. However, research from many different fields readily shows judgmental biases even in fields where high motivation to be accurate is expected, such as when grading students (Birkel, 1984a; Fiedler, Walther, Freytag, & Plessner, 2002; Rosenthal, 1991), judging gymnasts (e.g. Damisch, 2004; Plessner, 1997), deciding about penalties in soccer (Plessner & Betsch, 2001), or when making judgments in the courtroom (e.g. Downs & Lyons, 1991; Englich & Mussweiler, 2001) – this is even the case when judges are experts in their respective field (e.g. Downs & Lyons, 1991; Plessner, 1997; Plessner & Betsch, 2001) or when objective information is available to base the judgment on (e.g. Traut-Mattausch, Schulz-Hardt, Greitemeyer, & Frey, 2004).

When it comes to the grading of (final) exams (e.g., at high school or at university), it is very important that the judgments made are as accurate as possible. Therefore, research is necessary to find the possible sources of judgmental biases and to understand the underlying mechanisms of how people are making these judgments. This is also an additional goal underlying my work. I am especially interested in a source of errors that is interwoven with many judgment situations, especially when it comes to performance judgments such as grading exams of students. In these situations judges usually do not only make *one* judgment, but they usually have to make many similar judgments in a sequence. There are reasons to believe that sequential judgments may be prone to quite a number of different judgmental biases. Although my work focuses on the idea that prior judgments may influence subsequent judgments in a sequence, I will first give a brief overview of the different possible influences of sequences on judgments in the following section.

[4] One aspect of expertise is the experience with a certain task, namely that experts have more experience with a task than lay-people. This does not automatically imply that experts show also better performance at that task compared to lay-people.

1.3.2 How sequences affect judgments

In the following section I will present a brief overview of the literature on sequential judgment situations, showing the variety of potential biases that may influence judgments in these situation. Following this, I will present research especially focusing on the influence of prior judgments on subsequent ones. As explained above, one of my goals is to focus on more complex judgment situations as they are usually found in applied domains. Therefore, most of the examples presented in the following have the same focus.

One of the earliest examples of research on sequential judgments – at least in the field of social psychology[5] – may be the study of Sherif (1936) on how norms develop in small groups. Using an optical illusion known as the 'autokinetic effect', he let people judge the alleged movement of a point of light. In three sessions participants had to judge the movement in front of a group. Sherif found that the individual judgments, which were initially quite variable, later converged toward a common perception. In this example the judgment of other people influenced the following individual judgments; thus, this can be seen as an example of how group norms develop and may then influence individual judgments. A different focus on sequential judgments can be found in the area of performance judgments in sports, for example in the disciplines of gymnastics. When gymnasts of one team are judged in a sequence, the ones judged later usually get the better ratings (see Plessner, 1997, for an overview). This judgmental bias can be explained by an *expectancy effect* (see also Plessner, 1999). During a tournament teams in most cases decide on the performance order themselves, placing the better athletes at the end of the sequence. The judges know this fact: They expect the gymnasts performing later in a sequence to show a better performance, thus, leading to higher performance judgments independent of the actual performance. This example shows how expectancies about performances in a sequence may influence judgments.

In another study coming from the domain of sport psychology, Plessner and Betsch (2001) investigated whether there were sequential effects on penalty decisions of referees during a soccer match. Participants were in the role of a

[5] Work on sequential judgments can, of course, also be found in the domain of cognitive psychology. The interested reader can find an overview of most of the psychophysical models on sequence effects in a paper by Petzold and Haubensak (2001). Many of these models try to explain higher order sequential effects, i.e., effects of more than one stimulus or judgment preceding the current judgment of a stimulus.

referee. They were presented different ambiguous scenes of a soccer match, after which they had to judge these situations and decide on how to continue the game. They had to decide in particular for three successive foul scenes inside the penalty area whether to award a penalty or not. When participants had already awarded a penalty for one team, they were less likely to award a second penalty for the same team. However, they were more likely to award a penalty for the competing team. The authors explained this effect with the suggestion that referees followed, on the one hand, an equity-norm and, on the other hand, the unwritten rule that penalties should only be given rarely. Therefore, the decision criterion for referees shifted after they had awarded a penalty. Again, certain expectancies, rules, or norms – e.g., in this case about how many penalties should be awarded during a game – were guiding judgments and decisions in ambiguous situations.

Wedell and his colleagues (Wedell, Parducci, & Roman, 1989) investigated another unwritten rule – namely a fairness rule – that may guide the grading of exams. Based on the ideas of the *range-frequency theory* (Parducci, 1965) they wanted to see how participants would 'translate' scores of an exam to grades when instructed to do so as fairly as possible. The range-frequency model makes assumptions about category judgments, i.e., how people assign certain values[6] to given categories[7]. This is exactly the situation teachers are facing when they have to grade the written exams of a class or course. The theory assumes that two principles – the *range* and the *frequency principle* – guide these category judgments. The range principle makes assumptions about how judges divide the stimulus range to fit the categories of the judgment scale.

Following the range principle, judges will set a *psychological range* that is inferred from the stimulus range and will, then, set the sub-categories of the judgment scale in between. Since the psychological range is inferred from the range of stimuli to be judged (e.g., the exam scores of one course), the stimulus range sets the psychological range. In a course of only fairly intelligent students an average student would, therefore, be judged worse than in a course of fairly poor students.

The frequency principle additionally determines how frequently a judge uses different categories, i.e., how many stimulus values are assigned to each category. Following this principle, judges will assign to each category a fixed (in most cases an equal) number of stimuli values, e.g., the same number of As, Bs, Cs, Ds and Fs should be assigned for the exams of one course. Therefore, the judgments are influenced by the distribution or frequencies of the stimulus values. Additionally, if the

[6] E.g., the number of correct answers during an exam
[7] E.g., grades from A to F

frequencies of the stimulus values are not distributed equally, the range and the frequency principle are in conflict. According to the theory, a relative weighting of range and frequency tendencies solves this conflict. Usually both tendencies are weighed equally and the cut-off points of the categories lie in between the cut-off points suggested by the two principles. Wedell and colleagues could show in two experiments that participants did, indeed, follow these two principles (weighed about equally) when assigning grades to exam scores. Although they found no influence of the stimulus range on the psychological range as expected by the range principle[8], they could show that judgments of participants were influenced by the distribution of the exam scores as expected by the frequency principle. This example shows again how sequential judgments may be guided or biased by certain individual beliefs or unwritten rules.

A final example, which is also more related to performance judgments in exams, is the work by Betz (1974) on rhythmic oscillations. He found evidence for oscillation in the performance judgments during the sequence of oral examinations. This oscillation was independent of the performance shown by the tested students. Period and amplitude of the oscillation were actually caused by the number of examination sessions over one day. Betz, therefore, sees physiological factors, such as fatigue, hunger, and mental saturation, especially responsible for causing this oscillation. Since the influences of these physiological factors on judgments are not easy to prevent, he concludes that oral exams are no adequate way to test the abilities of students at all. This may only be the case when very few exams are taken in sequence. In this example the mere fact that judgments are made in a sequence is presented as the cause for the judgmental biases.

The examples presented so far show that there are many ways how judgments can be biased systematically in a sequential judgment situation. Developing or existing norms, expectations, and rules regarding the frequency distribution of judgment scores may influence judgments. Yet, there is also evidence for a more direct influence of one judgment on the one immediately following. This research is more directly related to the questions guiding my work and will be presented in the following section.

[8] Participants used the possible range of scores to determine the psychological range, not the range of the given stimulus sample.

1.3.3 How prior judgments influence subsequent judgments

The first question to be answered in my work asks whether there are systematic influences of prior judgments on subsequent ones in a sequential judgment situation. There are quite a number of studies showing that prior judgments influence following ones in complex judgment situations, an aspect on which I want to focus on with my work (e.g. K. R. Murphy, Balzer, Lockhart, & Eisenman, 1985; Smither, Reilly, & Buda, 1988; Sumer & Knight, 1996). In most of these studies, participants had to make evaluative judgments in a sequence about the performance of one (or two) person(s) in a job or task. In the work by Murphy and colleagues (K. R. Murphy et al., 1985, study 1), for example, participants had to view three videotaped lectures of the same lecturer and then evaluated these performances. The first two lectures were manipulated to show either a poor or a good performance; the third lecture showed an average performance and was the same for all participants. Murphy and colleagues found that the average performance ratings of the third lecture were significantly higher if preceded by the two poor lectures than if preceded by the two good lectures. Thus, they found a contrastive influence of prior performance judgments on subsequent performance ones.

Following this and other related studies, there seems to be ample evidence suggesting these effects of prior judgments on subsequent ones to be true in the area of grading exams. Yet, could these studies possibly have a conclusion regarding the direction of these effects as the second aspect of my first question, too? Even this question is addressed by Murphy and colleagues (K. R. Murphy et al., 1985). Although they provide evidence for contrast effects in their work, the authors assume that generally both, assimilation and contrast effects can possibly be expected. Moreover, based on the existing research “there may be little basis for predicting one type of effect over the other in a specific experiment or setting” (p. 75). This statement seems to be true: If one looks at the studies investigating the effects of prior judgments on subsequent ones in applied judgment situations, these studies have not yielded consistent results.

Some studies have found an assimilation effect, i.e., a bias toward the direction of the previous performance or judgment (e.g. Bazerman, Beekun, & Schoorman, 1982; Buda, 1986; K. R. Murphy, Gannett, Herr, & Chen, 1986a). Other studies have found a contrast effect, i.e., a bias away from the direction of the previous performance or judgment (e.g. K. R. Murphy et al., 1985; Schuh, 1978; Wexley, Yukl, Kovacs, & Sanders, 1972). There are also a number of studies trying

to resolve these empirical contradictions, which show both assimilation and contrast effects in performance judgments, depending on some manipulations of the judgment situation. As presented above, Murphy and colleagues (K. R. Murphy et al., 1985) found for example contrast effects of previous performance ratings on subsequent ones. In their studies, participants had to evaluate the performance of a lecturer giving three lectures, which were presented via video. The first two performances were either good or poor compared to the average third performance. Participants rated the average performance to be better after poor than after good performances. Murphy and colleagues explained the contrastive influence of prior performance ratings on subsequent ones with attention or encoding biases. In this case, the prior performance sets up expectations and, therefore, inconsistent elements in the subsequent performance capture more attention and will be encoded more richly. However, Murphy and colleagues (K. R. Murphy, Gannett, Herr, & Chen, 1986b) also found assimilation effects of subsequent performance ratings on previous ones. Participants had to watch and rate the same video tapes as described in the previous study. This time, yet, the first performance was average, followed by either poor or good performances. Additionally, participants had to judge the performance *after* having watched all three videos. This time, participants rated the average performance to be worse if it was followed by poor ones than after it was followed by good ones. Since attention and encoding cannot be influenced by expectations in this case, the authors explained this effect with memory biases. The overall impression of a judged person will be used to structure the performance information of the person judged stored in memory. Judges will, in this case, be biased in favor of recalling information consistent with the overall impression.

Another example is the work by Smither and colleagues (Smither et al., 1988), focusing on effects of prior performance information on ratings of subsequent performance. Participants had to rate the same videotapes that were used in the studies by Murphy and colleagues (1985, 1986). They had to judge the second, average performance either after having rated a (good or poor) first performance or after having received a (good or poor) performance rating. The authors could proof that the type or form of prior information (video vs. written rating) had an influence on performance ratings. Indirect information about prior performance given in a report from a credible source lead to assimilation effects whereas direct information about prior performance received via direct experience lead to contrast effects in the judgments of the present performance. The authors also showed that both, the time lag between the different performance ratings as well as the degree of extremity

between previous and present performance have an influence on the size of assimilation and contrast effects in the judgments of the present performance.

A final example is the work by Sumer and Knight (1996) on performance ratings of a secretary. Here, participants had to rate an average target performance after being exposed to either a good or a bad prior performance. The performances were presented through written scripts, which contained the secretary's job performance. The results of this study showed contrast effects in the judgments of the target if participants also had to judge the context performance, yet assimilation effects in the judgments of the target if participants were exposed only to the context performance without having to judge it. The authors explain this by referring to the *self-generated validity theory* (Feldman & Lynch, 1988). More precisely, they assume that raters will use prior performance evaluations as reference points for subsequent ratings, which leads to contrast effects when these performances depart from each other. However, when no prior performance evaluations are accessible, raters will use the memory of specific behaviors of the prior performance for building an evaluation of the target performance, which leads to assimilation effects.

To sum up, the results of the studies presented show that prior judgments have an influence on subsequent judgments in applied or natural judgment situations, i.e., at least in the field of work related performance judgments. As far as I know, there are no comparable studies in the field of performance judgments during examinations, apart from the work of Birkel (1978). In his study, 164 teachers from 39 different schools had to grade the performance of two pupils taking their final exams. The performance of the two students was presented on videotape. One of the pupils showed a better performance than the other one. The presentation-order of these two pupils was manipulated. Birkel found contrast effects in the performance judgments given by the teachers: The better pupil of the two was judged even better if in second position than if in first position; the worse pupil of the two was even judged worse if in second position than if in first position. Apart from this, Birkel also found an expectancy effect: If teachers received information about prior performances of the pupils to be judged, their judgments were influenced in direction of this information. So, this study gives additional evidence for effects of prior judgments on subsequent ones, as well as for the impact of expectancies on judgments.

Although there seems to be no question about the effects of prior performance judgments on subsequent ones, the direction of these effects is not easy to predict. Without a strong theoretical framework, it seems difficult to predict whether assimilation or contrast effects do occur in certain situations. As Murphy and

colleagues (1985) put it: “Although we expect that systematic differences in previous performance will affect evaluations of present performance, it is impossible, …, to confidently predict either assimilation or contrast effects” (pp. 76-77). To answer this question, I will now turn away from the area of applied research to a more fundamental one. As I have already outlined above, I regard comparisons between different performances as the driving force behind the effects of prior performance judgments on subsequent ones.

Therefore, I will take a look at the literature of social comparisons, a research field that originated from the *Social Comparison Theory* by Festinger (1954). I will then present the *Selective Accessibility Model* (SAM, Mussweiler, 2003a) as a framework that builds on the *Social Comparison Theory* in order to explain assimilation and contrast effects in human judgments. This model gives an answer to my questions, concerning both the direction and underlying processes of the effects of prior judgments on subsequent ones. Since many of the assumptions underlying the empirical part of my work (presented in chapter two and three) are derived from the SAM, this model will be presented in some more detail.

1.4 Comparisons cause assimilation and contrast effects in judgments

As I have outlined above, I expect that in the case of sequential judgments comparisons between prior and subsequent performances influence the judgment of the later target. Therefore, the focus of the following section will be on models concerned with social comparisons. First, the *Social Comparison Theory* of Leon Festinger (1954) – the father of research on social comparisons – will be presented. The research in line with this approach is, among others, concerned with the question of whether and how judgments may be influenced (or even biased) by comparison processes. After that I will outline the *Selective Accessibility Model* (SAM, Mussweiler, 2003a) as a model that predicts when to expect assimilation or contrast effects in judgments; the model also offers comparisons as a process underlying these effects. Finally, I will compare and distinguish the SAM from related and similar models.

To conclude, I believe that the use of theories and models concerned with social comparisons may be a helpful approach to answer the question about the direction of effects of prior judgments on subsequent ones and the question about the processes underlying these effects.

1.4.1 *Social Comparison Theory* – a selective review[9]

Festinger (1954) was the person who ’nvented’ the term *social comparison* and the first proposing a systematic theory about social comparisons with his *Theory of Social Comparison Processes* (or *Social Comparison Theory, SCT*). The theory is concerned with at least two basic questions: namely why (and when) and with whom do we compare ourselves. The original theory assumes that people possess a motivation to have a correct view of themselves[10]. Festinger derives from this assumption that we compare ourselves with others whenever there is no other possibility to accurately specify or evaluate our opinions and abilities. The theory also assumes that people will compare themselves with others that are similar to them on relevant dimensions. Although I assume that comparisons may have a biasing influence on judgments, the original *SCT* does not tell anything about how

[9] The following section is based on the work by Suls and Wheeler (2000)
[10] To be more precisely, of their opinions and abilities

comparisons may influence or bias the following judgments of our abilities or opinions. Festinger only states that people have an unidirectional drive upward for abilities, i.e., a motivation to become better on ability dimensions. However, the idea that the choice of a certain comparison standard has important consequences and may influence our judgments (and also feelings) was developed later.

Following theories expanded the SCT by adding a self-enhancement motive to the self-evaluation motive (e.g. Hakmiller, 1966; Thornton & Arrowood, 1966; Wills, 1981). This self-enhancement motive provides people with a reason for comparing themselves in a way that will bolster their self-esteem. One very prominent approach concerned with this self-enhancement motive is probably the work by Wills (1981), focusing on the concept of downward comparisons[11]. It is his main idea that people can increase their well being by comparing themselves with others who are worse off on relevant dimensions. This idea was fostered by the work of Wood and colleagues (Wood, Taylor, & Lichtman, 1985), showing that even cancer patients were spontaneously comparing their own conditions with patients who were less fortunate, thus, using downward comparisons to increase their subjective well-being.

Underlying the assumed self-enhancement motive and the use of downward comparisons is the assumption that people will strategically look for comparison standards in order to influence their self-judgments. Thus, in these theories a strategic judgmental bias (or influence) as a consequence of comparisons is a necessary pre-condition to increase self-esteem or subjective well-being. Thus, the choice of a comparison standard has an important influence on the outcome of comparisons.

Yet, in which direction do comparisons influence our judgments? Wills (1981) assumes in his concept of downward comparisons that comparisons have a contrastive influence on our judgments. Therefore, if we compare ourselves with people who are better on a certain dimension, we will end up with a worse judgment of ourselves than if we compare ourselves with people who are worse on that dimension. This one-sided view had already been questioned before by the work of Thornton and Arrowood (1966). The authors also posited a self-enhancement motive, nevertheless assuming that people compare themselves with others that are *better* on a certain dimension in order to increase their self-evaluation. In this case, judges get a better view of themselves by focusing on the assumed similarity to these better comparison standards, judging themselves as being nearly as good. The authors, thus, assume an assimilative influence of comparisons on our judgments.

[11] This idea was originally introduced by (Hakmiller, 1966).

Taken together, there seem to be different consequences of comparisons independent of the selected comparison standard. This idea can also be found in the *Self-Evaluation Maintenance Model* (SEM) by Tesser (1988). This model focuses solely on the motivation to maintain or enhance self-evaluation. Here, the performance or attributes of other persons may either have an assimilative or a contrastive influence on our self-evaluations. If people follow a *comparison process*, this will have contrastive consequences; whereas if people follow a *reflection process*, this will have an assimilative consequence ("basking in the glory of others").

More recent approaches integrate more explicitly the idea that social comparisons may have assimilative as well as contrastive consequences. Buunk and colleagues (Buunk, Collins, Taylor, & Van Yperen, 1990) state that comparisons can produce positive or negative feelings independent of the direction of the comparisons. Also, Taylor and Lobel (1989) argue that both upward and downward comparisons may be self-enhancing, depending on the type of activity people follow. A desire for information about others may have contrastive consequences, for example, whereas a desire to affiliate with others may have assimilative consequences. A third recent approach is the Selective-Accessibility Model (SAM) by Mussweiler (2003a). This model also assumes that comparisons may have assimilative as well as contrastive consequences on judgments, depending on the comparison focus people employ. Apart from that, this model is moreover an integrative approach, combining comparison theory with the social cognition literature on assimilation and contrast effects.

To conclude, judgments that are based on comparisons may be influenced (or biased) by these comparisons. On the one hand, the selection of the comparison standard has an important influence on the judgment; on the other hand, the direction of this influence may possibly go in two directions and can either have assimilative or contrastive consequences. Therefore, I believe that comparisons may play an important role for sequential judgment situations, too, and may be useful to explain the effects of prior judgments on subsequent ones. In the following section, I will present the SAM in some more detail, a recent model on which many of the assumptions underlying the empirical part of my work are based.

1.4.2 The *Selective-Accessibility Model* (SAM)

As I have mentioned above, some of the basic hypotheses of my empirical work are based on the assumptions of the SAM (Mussweiler, 2003a), this model will be introduced in some detail in the following section. This approach seems, for many reasons, well suited to answer the questions regarding the existence and direction of systematic effects of prior judgments on subsequent ones and the processes underlying these effects. The SAM is a recent model that builds on and integrates assumptions of earlier models, e.g., the *Inclusion/Exclusion Model* (Schwarz & Bless, 1992a). The model assumes that comparisons are an important variable or process influencing any given judgment and it furthermore allows precise predictions on the direction of these influences. In the situation of sequential performance judgments, I also believe that the comparison of prior with subsequent performances is an especially important variable or process influencing the judgments. Finally, the model focuses particularly on the active process of selective information search or activation. In the situation I focus on with my work – the case of sequential judgment when grading (oral) exams – a judge is not simply passive, receiving information, but actively searching for it. Therefore, I believe that the SAM is a model well suited to answer the questions I raised, especially for the situation of making judgments in (oral) exams. I will now give a detailed presentation of this model, including an overview of the related empirical work. I will, finally, compare the SAM with (and distinguish it from) similar approaches explaining assimilation and contrast effects in judgments.

Mussweiler (2003a) introduces his SAM with the following statement: “Human judgment is comparative in nature.” (p. 472). With this short statement he already outlines two main assumptions of his model. First, there are no absolute ratings when it comes to social judgments[12]. Second, there is a process of comparison of the target to be judged with a given standard preceding any (social) judgment. The main goal of the model is to explain the psychological mechanisms underlying assimilation and contrast effects in social judgments. By focusing on basic cognitive processes that guide comparisons and social judgments, the model offers very precise predictions about when assimilation and contrast effects in judgments should occur.

[12] Mussweiler (2003a) distinguishes judgments and comparisons made in the social domain – also called informational or noetic – from experiential judgments and comparisons that are directly based on sensory input (cf. Strack, 1992)

Thus, this model can be understood as an “unifying theoretical model that is able to integrate the diverse consequences of comparisons” (Mussweiler, 2003a, p. 473).

The basic idea of the model is very simple: There is no judgment without comparison. The model offers no proof for this assumption, although many examples are offered that show the prominent role of comparisons for person perception (Herr, 1986), decision making (Kahneman & Miller, 2002), and judgments of the self (Festinger, 1954). Apart from this exemplary evidence, no mandatory evidence on a theoretical level is offered for this statement, which means that this statement has to be taken for true like an axiom. Any judgment, therefore, depends on the preceding comparisons. The SAM focuses mainly on the stage of comparison, assuming that the basis for assimilation and contrast effects can be found exactly there. To be able to judge the importance of this comparison process and to understand how it influences judgments, a general overview of the process of judgments seems appropriate and will be presented first. After this overview, a closer look at the central part of human judgments according to the SAM – namely, the “selective accessibility mechanism” – will follow.

According to the axiom that there are no judgments without comparisons, the process of comparative evaluation of a target can be separated into three stages: First, the s*election* of a comparison standard and of comparison features, second, the realization of the *comparison* between target and standard, and third, the *evaluation* or judgment of the target. As explained above, the SAM focuses on the second stage, the process of comparison. The assumptions concerning the other two stages are derived from the existing literature.

The selection stage. Whenever one wants to make a comparison, one obviously needs two objects or instances to compare each other with. Using the terms of comparison research, for a comparison I need a *comparison standard* to compare the *target* (later to be judged) with. It seems important to find a standard that is relevant to the judgment task. Yet, in many situations there is more than one relevant standard available that could be used to compare the target with. The situation of grading an exam can serve as an example for this idea. As a judge, one can think of quite a few relevant standards to compare the performance of the target exam with. One could use various objective (e.g., the number of correct answers for a specific grade) as well as social standards (e.g., the performance of the previously judged exam).

Now there is still the question, what makes one possible standard more relevant than the other? At least three basic principles can be derived from the

literature: *Conversational inferences*, *accessibility*, and *normative concerns*. First, following the cooperative principle built on the conversational maxims as suggested by Grice (2002), a judge should assume a given standard to be relevant if it is explicitly or implicitly offered by a communicational partner (e.g., the experimenter in an experimental setting). Therefore, conversational inferences may make a given standard appear to be more relevant than others (Schwarz, 1994). Second, a standard may be chosen because it is highly accessible either due to being salient in a given situation or due to the accessibility in memory (e.g., Herr, 1986). Third, normative concerns may also guide the selection of a relevant or diagnostic standard. Especially similarity of target and standard seems to play an important role when it comes to comparability of standard and target (Aarts & Dijksterhuis, 2002; Festinger, 1954; Smith & Zárate, 1992). When it comes to judge a target, it seems advisable to make sure that one is not comparing apples with pears.

Considering these three basic principles guiding standard selection, what could be expected for the performance judgment example? Conversational inferences should only be influential if there is a communicational partner available when making the judgment. This seems not to be very likely in the case of grading exams[13]. Accessibility seems to be the more important variable. I would expect the concepts of primacy, recency, and peak to influence accessibility in sequential judgments. Therefore, the first exam in a sequence, the one prior to the present exam, and exceptionally good or bad exams should be most accessible as possible standards. Finally, it will be important to determine whether it seems normatively correct to compare performances in exams, in general. On the one hand, it might very well be that a 'fairness norm' prohibits the comparison of performances in exams. On the other hand, it may also seem unfair or not diagnostic to compare a very good with a mediocre or bad performance. So, it may well be that fair examiners would rather choose an objective or abstract standard of comparison.

Apart from selecting a standard of comparison, a judge also has to determine what features are relevant for the comparison. In the example of grading exams, this may be the number of correct and incorrect answers, the quality of the language used, and the number of spelling mistakes. Both, the selection of standard and target do have an important impact on the outcome of the comparison and the subsequent judgment. One could say that "standard selection and featural focus set the informational stage for comparison consequences" (Mussweiler, 2003a, p. 474) and, thus, for the final judgment. When I have to judge the temperature in a room, it

[13] An exception would be an assessor who could offer a possible standard during an oral exam.

makes a difference whether I just came from outside where a snowstorm was ravaging, or if I just stepped outside of a sauna. In a similar same way, a student taking an exam appears in a different light when compared to the best or the worst student of his year.

The comparison stage. Although the first stage already has an influence on the outcome of the judgment process, this second, comparison stage is of at least equal importance according to the SAM. During this stage it is decided whether assimilation or contrast effects will occur in the final judgment. The information that is activated during this stage will be most accessible at the final stage of evaluation and the evaluative judgment of the target will, therefore, be based on this information. The SAM assumes that information during the comparison will be activated selectively. This means that judges do not randomly draw samples of the pool of information available. They rather follow a hypothesis or expectation that they are testing during the comparison. Testing a hypothesis makes judges focus selectively on certain bits of information, which will be used as a basis for the following judgment. This *selective accessibility mechanism* will be explained in some more detail below. It is important to note that neither the selection of the standard nor the comparison of target and standard have to be a deliberative or conscious process (e.g., Dehaene et al., 1998; Dunning & Hayes, 1996; Gilbert, Giesler, & Morris, 1995).

The evaluation stage. In this final stage, the information collected about the target in the prior stages will be integrated into an overall evaluation. As explained before, the SAM assumes that, in this stage of target evaluation, the information, which is most accessible, will be used to build the evaluative judgment on. This idea is derived from more general principles guiding acquisition and use of knowledge (see Higgins, 1996). Therefore, the evaluation of the target will likely be consistent with the information activated during the comparison stage. This means that the selectivity of information search during the comparison phase will consequently influence the judgment of the target. Yet, why is information search selective? Moreover, when does a selective search lead to assimilation and when to contrast effects? These questions are explained by the *selective accessibility mechanism* that will be presented in the following section (see also figure 1.1).

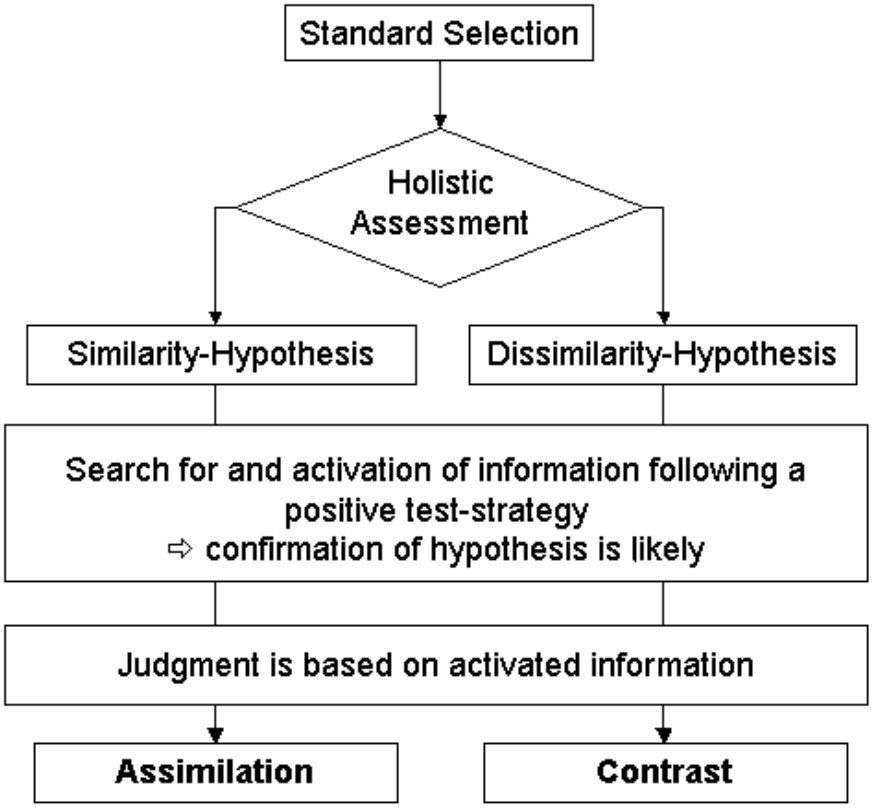


Figure 1.1. The Selective-Accessibility Model

The selective accessibility mechanism. This mechanism is the central part of the SAM. It tries to explain the process underlying assimilation and contrast effects in judgments. As mentioned above, this mechanism influences the comparison of target and standard prior to the judgment of the target. It is assumed that judges first come up with a single focal hypothesis about the relation of target and standard (see Sanbonmatsu, Posavac, Kardes, & Mantel, 1998; Trope & Liberman, 1996). This makes it possible to distinguish relevant from irrelevant information and, therefore, the comparison task easier to accomplish. The model states that judges will always focus on one of the two possible hypotheses. Judges can assume that target and standard are either similar (similarity-hypothesis) or dissimilar (dissimilarity-hypothesis). Although not explicitly stated, one can easily translate these hypotheses into more precise expectations by taking into account the chosen standard of comparison. This can be demonstrated by looking at the example of grading exams. If a judge compares the target exam with a very good (or high) standard and focuses on similarities, this may result in the expectation that the target exam will be (as) good (as the standard). However, if a judge using the same good comparison

standard focuses on dissimilarities, this may result in the expectation that the target will not be (as) good (as the standard). Still, how does a judge decide, which of the two possible hypotheses to use? The SAM postulates that judges will first engage in a *quick holistic assessment* of target and standard. During this assessment, a small number of salient features is used to come up with a first similarity judgment and the resulting hypothesis. As stated above, similarity between target and standard plays an important role during the stage of standard selection. Only a standard that is similar enough may be chosen to be compared with the target. Mussweiler derives from this assumption that judges will, in most cases, come up with a similarity-hypothesis. It is important to note that this first similarity judgment is supposed to be too general to solely base the target evaluation on.

Once judges have developed a first idea that target and standard are either similar or dissimilar, they will test this hypothesis. The SAM reverts to the findings in the literature on hypothesis testing and states that an active (though not necessarily deliberative or consciously aware) process of hypothesis testing will follow. According to the SAM, people will test their hypothesis by focusing on hypothesis-consistent evidence (cf. Trope & Liberman, 1996), i.e., the hypothesis under test will influence the test-strategy. This may result in either a confirmatory test-strategy (e.g. Snyder & Swann, 1978) or a positive test-strategy (Klayman & Ha, 1987). In the former case, motivational reasons are seen as the causing factor and the judgment will always be influenced in the expected direction. In the latter case, the test-strategy is influenced by cognitive factors and the interaction of test-strategy with the true allocation of the tested variable will determine how the judgment will be influenced. The SAM assumes that “judges selectively generate information that is consistent with the focal hypothesis of the comparison” (Mussweiler, 2003a, p. 475). This may be caused either by a confirmatory or a positive test-strategy. In the latter case, it is important whether hypothesis-consistent information is found; only then the hypothesis will also be confirmed.

The idea that expectations or hypotheses will guide our view of the world is widespread in the area of social cognition. Once an idea is formed or activated in our head, we will see the world in a different light, i.e., we will perceive, activate and interpret information in a way that is consistent with our idea, resulting in simple and well-structured impressions of our world (cf. Fiske & Taylor, 1991). This is true for any kind of idea or knowledge structure such as attitudes (Fazio & Towles-Schwen, 1999), stereotypes (Bodenhausen & Macrae, 1998), or hypotheses (Trope & Liberman, 1996). For the case of hypotheses, this means that when people test their hypotheses, they may (automatically) activate and retrieve evidence from memory

(Koehler, 1991; Snyder & Uranowitz, 1978) and may encode ambiguous information in line with the hypothesis being tested (Higgins & King, 1981; Srull & Wyer, 1989). Thus, by looking for consistent evidence, people may end up thinking they actually found evidence for the hypothesis and judge this hypothesis to be true. For the SAM, this means that judges will be likely to find and activate more hypothesis-consistent than hypothesis-inconsistent information, i.e., it is likely that the hypothesis will be confirmed. Since the final evaluation of the target will be based on previously activated information, the judgment will also be in line with the hypothesis or expectation of the judges. Thus, when judges focus in similarities, they will activate information consistent with this hypothesis and their judgments will show assimilation effects. However, when judges focus on dissimilarities, they will activate information consistent with that hypothesis and their judgments will show contrast effects. Consider the example of grading exams: If judges expect an exam to be good, they will focus on information that is consistent with that idea (and may even interpret ambiguous information accordingly) and will, therefore, finally grade that exam better than if they had expected the same exam to be bad.

The idea of the selective accessibility mechanism is very closely related to what recent models of stereotyping propose (e.g., Bodenhausen & Macrae, 1998; Devine, 1989; Fiske, Lin, & Neuberg, 1999; Kunda & Spencer, 2003; Lepore & Brown, 1997). In their general model of the influence of stereotypes on interpersonal impressions, judgments, and behaviors, Bodenhausen and Macrae (1998) say that in person judgments a process of categorization and stereotype activation is started first. This can be compared with the quick holistic assessment in the SAM. It depends on factors such as (contextual or habitual) salience, recency, chronic use, and perceiver's momentary goals, which category will be activated. These factors may also be important for predicting whether a judge will follow a similarity- or dissimilarity-hypothesis, according to the SAM. Once a certain category and associated stereotypes are activated, these will facilitate the activation and processing of consistent data. Moreover, this activation of a certain category and associated stereotypes will inhibit the activation and processing of inconsistent data. Additionally, the activated stereotypes will influence the interpretation of activated or sought data, assimilating it (i.e., bringing it in line with) to the activated stereotypes. This, again, is consistent with the idea of the SAM that once a hypothesis or expectancy is activated, it will influence the activation and interpretation of information when testing the hypothesis.

In the SAM, the *quick holistic assessment* seems to play a very important role. Since it is here that the judge decides to focus on either similarities or dissimilarities

leading to assimilation or contrast effects in the final judgment. Yet, who makes this decision or how is this decision made? Here, the model is not very precise and this makes it, on the one hand, flexible enough to explain many findings on assimilation and contrast effects reported in the literature, but, on the other hand, makes it difficult to come up with precise predictions on when to expect assimilation and contrast effects. “Any factor that influences the nature of the initial hypothesis is likely to have an effect on comparison consequences” (Mussweiler, 2003a, p.486). Any factor that is salient in a given situation, may influence the outcome of the quick holistic judgment, such as category membership (Mussweiler & Bodenhausen, 2002), extremity of the comparison standard (Herr, 1986), ambiguity of the target (Stapel, Koomen, & van der Pligt, 1996), or unique (Houston, Sherman, & Baker, 1989) and shared features (Gati & Tversky, 1987; see also Tversky & Gati, 1978). Again, even if the model can explain the processes underlying assimilation and contrast effects – and there is large number of empirical findings supporting the assumptions of the model (see below) –, it does not easily allow precise predictions about when assimilation and contrast effects will occur.

It is important to note that, as mentioned above, only standards that are fairly similar to the target are used to compare the target with. Therefore, similarity testing and assimilation effects seem to be the default outcome of comparison based judgments. Thus, “the structural requirements of the initial similarity assessment … gear judges toward similarity testing” (Mussweiler, 2003a, p. 479).

Before I will try to distinguish the SAM from other, similar approaches explaining assimilation and contrast effects in judgment, I will first give a brief overview of the empirical evidence supporting the assumptions of the SAM. Most of the empirical support for the selective accessibility mechanism that is proposed by the SAM comes from work examining comparison processes using the paradigms of ‘social comparison’ and of ‘judgmental anchoring’. In both paradigms, participants are first ask for a relative judgment, i.e., a judgment of the target on a certain dimension in comparison to a given standard (e.g., “Are you more or less famous than Batman?”). After this relative judgment, participants are asked to make an absolute judgment on the same dimension (e.g., “How famous are you?”). The main difference between the two paradigms is that in most social comparison paradigms the target is the judge and the standard another person, whereas in most anchoring paradigm the target may be anything or any person (e.g., the river ‘Elbe’ or ‘Gandhi’) and the standard is a number (e.g., ‘2000 km long’ or ‘150 years old’). By focusing only on these two paradigms, the impact of the empirical evidence or the external validity is

restrained. The paradigm of social comparison solely focuses on judgments of the self. The self is a very special target that can be distinguished form many other targets for many reasons, such as the amount, complexity and structure of available and accessible knowledge about the target (see Fiske & Taylor, 1991). “Despite the general similarity of the ways people learn about themselves and others, self-knowledge is richer and more detailed than knowledge about others. People can observe themselves in more situations and have better access to private thoughts and feelings.” (Smith & Mackie, 1995, p. 119). Judgments about the self may be guided by motives that differ from those guiding other judgments (e.g., self-enhancement motive). Moreover, the search for information preceding the judgment of the self is a memory-based process (see Hastie & Park, 1986, for a distinction of memory-based vs. on-line judgments). In most studies using the judgmental anchoring paradigm, judges are explicitly instructed to first compare the target to be judged with a (numerical) standard. Additionally, in most of these studies only fairly irrelevant judgments were to be made, such as the length of a river or the price of a car. The judgment tasks used in both paradigms are, in most cases, rather simple ones.

This critique concerns the ecological or external validity, a very important point in my opinion. The importance or relevance of the SAM can partly be determined by its explanatory or predictive power in more complex, applied judgment situations that usually are situated within a rich and less structured environment. There is not much empirical evidence in such judgment situations, with some noteworthy exceptions. First, the work by Englich and colleagues on anchoring effects in the courtroom (e.g. Englich & Mussweiler, 2001) and second, the work by Damisch (2004) who directly tested assumptions of the SAM regarding assimilation and contrast effects by examining professional judges evaluating gymnasts. These examples give reason to believe that the assumptions of the SAM may also be useful for more complex judgment situations, like the one used in my studies.

Let me now focus on the existing work testing the assumptions of the SAM. Following the selective accessibility mechanism, judges first have to come up with a quick holistic assessment to determine the similarity of target and standard. This quick holistic assessment will be based on a consideration of a small number of salient features, such as category membership. In line with this, Mussweiler and Bodenhausen (2002) could show that, following a comparison of oneself as a target with an extra categorical standard (e.g., a male judge comparing himself with a female standard), information about the category membership of the target is more

accessible. However, if target and standard belong to the same category, specific individuating knowledge about the self will be more accessible.

The SAM assumes further that this first holistic assessment of similarity is not enough to base the final judgment of the target on. To show the importance of information that is activated during the comparison process, which follows the quick holistic assessment, Mussweiler and Strack (1999) manipulated the amount of time judges had when comparing target and standard in an anchoring paradigm. Half of the participants made this comparison without time constraints, the other half was only given five seconds for this task. The assumption was that five seconds should be enough time to accomplish the quick holistic assessment of similarity, but not enough time to generate the knowledge necessary to subsequently evaluate the target. If now the final evaluation of the target was simply build on the quick holistic assessment of similarity, there should be no difference between the two groups regarding the time needed to come up with the final judgment of the target. However, the authors found a difference: Participants with time constraints needed more time for the final judgment than participants without time constraints. This means that the quick holistic assessment of similarity activates not enough information to come up with a final judgment.

The SAM next proposes a comparison process that is guided by a hypothesis. Therefore, this process should be an active process of hypothesis testing that could be distinguished from more passive processes of knowledge activation by excitation transfer. Although such passive processes may play a role in comparison processes, there are important differences between these two processes. First, knowledge activation via spreading activation increases the accessibility of knowledge structures or concepts independent of the target to be judged. Second, passive knowledge activation increases the accessibility of linked knowledge structures independent of the direction of the comparison (e.g., the nature of the tested hypothesis). There are quite a few studies proving that comparison processes in social judgment do activate knowledge that is specific to the target (e.g. Mussweiler & Strack, 2000a; Mussweiler & Strack, 2000b) as well as studies proving that differential knowledge activation depending on the direction of the comparison or the nature of the tested hypothesis (e.g. Mussweiler & Strack, 1999).

Apart from this evidence proving that not only passive processes of knowledge activation play a role in comparisons, there seem to be no studies focusing explicitly on the process of hypothesis testing. In most studies examining the general processes of hypothesis testing, the focus is on the queries people make, i.e., the amount and kind of information people search for by asking certain kinds of

questions (see e.g., Devine, Hirt, & Gehrke, 1990; Skov & Sherman, 1986; Snyder & Swann, 1978; Trope & Liberman, 1996; Trope & Thompson, 1997; Zuckerman, Knee, Hodgins, & Miyake, 1995). Although there are some studies showing that hypothesis-consistent information is activated after the target is judged (by using a type of lexical decision task e.g., Mussweiler & Strack, 2000b, 2000c) and although there also exists one study using a thought-listing procedure to see what concepts are activated during the comparison process (Mussweiler & Strack, 1999, study 4), it is important in my opinion to examine the assumed process of hypothesis testing more directly by focusing on the amount and kind of information sought. Especially, since the SAM directly derives its ideas on the stage of hypothesis testing from the general literature mentioned above (e.g., Snyder & Swann, 1978; Trope & Liberman, 1996).

Finally, some of the best empirical support for the SAM and the selective accessibility mechanism comes from a study directly manipulating the hypothesis of participants by letting them either focus on similarities or differences in an unrelated comparison task preceding the relevant judgment task (Mussweiler, 2001b). Here, participants showed assimilation effects in their judgments when primed to focus on similarities and contrast effects when primed to focus on dissimilarities.

1.4.3 A comparison model in comparison with other models

The SAM is by far not the only model that tries to explain assimilation and contrast effects in judgments. (see Ford & Thompson, 2000; Petzhold & Haubensak, 2001; Stapel & Koomen, 2001b, for an overview). In the following section, I will describe related models that make similar assumptions regarding assimilation and contrast effects in judgments and I will try to distinguish these models from the SAM.

The Inclusion/Exclusion Model (Schwarz & Bless, 1992a). One of the two main assumptions of this model is that accessible information at the time of judgment influences the judgment of a target. This seems compatible with the assumptions of the SAM. Yet, while the model by Mussweiler (2003a) focuses on the kind of information that becomes accessible during comparison, the *Inclusion/Exclusion Model* focuses on how this accessible information will be used when forming a judgment. Therefore, a second main assumption is proposed, namely that the categorization of accessible information will determine whether this information is either included in or excluded from the target category, thus, resulting in assimilation

or contrast effects, respectively, in the judgment of the target. The *Inclusion/Exclusion Model* states that to judge a target one first has to build some cognitive representation of it. In addition, one has to determine some standard of comparison to evaluate the stimulus[14]. Thus, this model also seems to suggest that human judgment is comparative in nature. Unfortunately, the model does not further elaborate on this point and comparisons seem to play an important role only for contrast effects, but not assimilation effects (see below). When building the cognitive representation of a target, judges either add accessible information to or subtract it from the representation of the target. Adding information to the representation of the target will lead to assimilation effects, i.e., the information will be included in the judgment of the target. This can only be empirically observed when a given piece of information differs from the other accessible pieces of information.

Imagine the example of grading the performance of a student taking an oral exam. Maybe this student gave a poor answer to the first question of the exam. A judge will include or assimilate this piece of information to the representation of the overall performance of the student. This may let the performance look worse if the other answers of the student were better as the first one. However, it will not change the performance judgment if the student answered all of the questions poorly. If a piece of information is not included into the representation of the target, this will lead to a subtraction effect and subsequently to contrast effects in the judgment. Thus, even the judge knows how well a student performed during his previous exams, he may not want to use this information when judging the present performance[15]. These contrast effects can also only be observed if the subtracted information differs from the rest of the accessible information.

The question remains, how it is decided, if a given piece of information is added to or subtracted from, included into or excluded from the representation of a target. According to the *Inclusion/Exclusion Model*, this is decided during the stage of categorization. If some information belongs to the same category as the target to be judged, this will lead to inclusion and hence to assimilation effects; if some information belongs to a different category as the target, this will lead to exclusion and hence to contrast effects. In the example stated above, teachers will probably include all of the answers of a student given during one exam into the performance judgment of that student, but they will exclude the performance of a prior student

[14] Context influences on judgments are only expected when judgments are formed on the spot and not when previously formed judgments are retrieved from memory (see Strack & Martin, 1987).
[15] Although some teachers feel that it is appropriate to include this kind of information into the performance judgment (e.g., “I know this student from my class and usually this student performs better”)

taking the same exam, unless teachers will categorize the two students as belonging to the same category. Variables like the width of the target category (e.g., Schwarz & Bless, 1992b), the extremity (e.g., Herr, Sherman, & Fazio, 1983), or, more general, the representativeness (e.g., temporal distance, Strack, Schwarz, & Gschneidinger, 1985) of accessible information for the target category will influence whether a piece of information is included into or excluded from the representation of a target.

This assumption seems familiar. One could translate this stage of categorization into the stage of quick holistic assessment in the SAM by exchanging ‘belonging to same category’ with ‘similarity judgment’ and ‘belonging to different categories’ with ‘dissimilarity judgment’. However, the major difference between these two models is what follows from this categorization / quick holistic assessment. The *Inclusion/Exclusion Model* assumes that because of this categorization, already accessible information will be used when judging a target. In contrast to this, the SAM assumes that the outcome of the quick holistic assessment directly influences the activation or accessibility of information.

A second difference is the role of comparison processes in judgments. Both models assume that all (social) judgments are comparative in nature. However, whereas the SAM assumes that comparisons can lead to both, assimilation and contrast effects, the *Inclusion/Exclusion Model* assumes that a target can only be compared to a standard that is excluded from the target category, i.e., comparisons will always lead to contrast effects. The *Inclusion/Exclusion Model* offers, therefore, comparisons as a second possible process underlying contrast effects in judgments apart from subtracting effects. Apart from constructing a cognitive representation of the target a judge also has to construct a comparison standard. Any kind of information that is excluded from the target category may be used (in an additive fashion) for constructing a comparison standard. The model implicitly assumes that comparisons always lead to contrast effects in judgments. The two possible processes underlying contrast effects can be distinguished regarding to the model. The subtraction of information should only affect the judgment of the target category from which the information is excluded. When information is used to construct a comparison standard this should also affect the evaluation of related stimuli to which the standard may be relevant. In the studies reported by Strack and Bless (1992a), they find evidence as well for the subtraction effect as for the comparison effect.

Finally, a third explanation for contrast effects is offered. Instead of using excluded information for constructing a comparison standard, this information may also be used to anchor the response scale (e.g., Ostrom & Upshaw, 1968). The use of some information as a reference point or anchor instead as a comparison standard

is also offered by Mussweiler (2003a) as a possible alternative explanation for contrast effects. Yet, these two processes are difficult to distinguish on the empirical level (see Eiser, 1990). However, they do not have to be mutually exclusive and may both operate in parallel, thus, jointly contributing to contrast effects.

Summing up, the *Inclusion/Exclusion Model* makes many similar assumptions as the SAM about when to expect assimilation or contrast effects (e.g., if target and standard are similar (dissimilar) this should lead to assimilation (contrast) effects in judgments of the target). Yet, the two models offer different underlying processes that lead to assimilation and contrast effects. Most importantly, the *Inclusion/Exclusion Model* sees the differential use of activated information as the cause for assimilation and contrast effects whereas the SAM sees the differential (or selective) activation of information as the cause.

The Interpretation/Comparison Model (Stapel & Koomen, 2001a). This model is anchored in the research tradition of impression formation. Starting with the work of Higgins, Rholes and Johnson (1977) and Srull and Wyer (1979), it has been shown that prior activation or priming of (category) information influences the interpretation of ambiguous behaviors when forming an impression of a person[16]. Activated information may, therefore, again lead to assimilation and contrast effects in the judgment of a person. Stapel and Koomen (2001a) say that to form an impression of a person, one first has to encode and identify the observed behaviors and second has to combine these behaviors into a judgment on a given dimension. This idea is derived from earlier theoretical work, e.g., a model by Trope (1986) distinguishing the stages of identification and inference within the impression formation process. The *Interpretation/Comparison Model* follows the basic assumption of the model by Trope that the outcome of both interpretation and judgment processes is determined by accessible information. The model further states that accessible information will have an assimilation effect during behavior interpretation, but a contrastive effect during the construction of the judgment. During the interpretation stage, accessible information will be used as an interpretive frame, therefore, interpretative assimilation should occur. During the judgment stage, accessible information will be used as a comparison standard, therefore, comparative contrast should occur.

[16] Although its seems to be tradition to name the person to be judged ‘Donald’, this is true, even if the person judged has a different name. “Despite the concerns of a generation of graduate students, we can state with some confidence that naming the target ‘Donald’ is *not* a necessary step to obtaining priming effects” (Bargh & Chartrand, 2000, p.11)

Think again of the exam situation as an example. Teachers may use the information how well formulated the answers of a student were (form) as a cue to interpret the quality or correctness (content) of the answers and, thus, the overall performance of the student. In this case, the performance judgment based on the content judgment should be assimilated toward the form judgment. This may of course only happen, if the performance of the student is ambiguous enough to leave room for interpretation. However, it may be that the way the answers are presented may be used as a standard of comparison, thus, resulting in contrast effects. For example, the content of the answers may look especially good in comparison to the way they were presented. In this case no ambiguity is needed.

Summing up, accessible information may again be the cause for both, assimilation and contrast effects, depending on the stage it is used during the process of impression formation. Since the same kind of information may be used at both stages of the impression formation process, it is important to understand what determines during which stage a given piece of information is used. The *Interpretation/Comparison Model* assumes that this may be determined by characteristics of the accessible information itself or by the activated goals of the person forming an impression. Examples for the characteristics of accessible information are extremity (e.g., Herr, 1986; Manis, Nelson, & Shedler, 1988) and abstractness of this information. Regarding the latter characteristic, Stapel and colleagues demonstrated that more abstract information (e.g., traits) is used for interpretative purposes, thus, producing assimilation effects, whereas more concrete information (e.g., exemplars) is used as comparison standard, thus, producing contrast effects (e.g., Stapel et al., 1996; Stapel, Koomen, & Zeelenberg, 1998).

One distinguishing feature of the *Interpretation/Comparison Model* is the focus on goals as a determinant feature for assimilation and contrast effects. The model assumes that judges may either be in an interpretative or a comparison mindset[17]. This mindset determines whether a given piece of information is used either during the interpretation or during the judgment stage, independent from impact of the characteristic of this information. Stapel and Koomen (2001a) demonstrated this in a series of studies where they manipulated the respective mindset using a (implicit) priming procedure.

By comparing the *Inclusion/Exclusion Model* (Schwarz & Bless, 1992b) with the *Interpretation/Comparison Model* (Stapel & Koomen, 2001a) described in this section, one may notice that both models have more in common than just the '/' in

[17] Although one can think of the possibility that the same kind of information may be used at both stages, the authors do not address this issue but stay with a maybe artificial dichotomy.

their names. It seems again that both models make fairly similar assumptions[18]. One the one hand, accessible information may be used to 'build a cognitive representation of a target' according to the *Inclusion/Exclusion Model* or as an 'interpretation frame' to understand the target behavior according to the *Interpretation/Comparison Model*. On the other hand, the same kind of information may be used as a comparison standard (or to construct a comparison standard) according to both models. Of course, the *Inclusion/Exclusion Model* offers two additional explanations for contrast effects (subtraction effects and anchoring effects), but the main difference between these two models is that the *Interpretation/Comparison Model* makes more explicit assumptions concerning the (time flow of the) different stages of the judgment or impression formation process.

The differences to the SAM (Mussweiler, 2003a) are, therefore, the same as for the *Inclusion/Exclusion Model* (see above). The main difference is that the SAM is concerned with the *selective activation of information*, whereas the Interpretation/Comparison Model makes assumptions about the *selective use of accessible information*. There is also an interesting assumption by the *Interpretation/Comparison Model* that differs from the two other models. Stapel and Koomen (2001a) state that the context-target similarity may also determine assimilation and contrast effects. They assume that if target and context information are categorized as being similar, comparison contrast will occur, whereas when both are categorized as being different from each other, assimilation will occur (see also Kahneman & Miller, 1986). This assumption contradicts the assumptions of the two other models presented before. These models both assume that similarity leads to assimilation rather than contrast effects. Kirsten Ruys (2004) tries to overcome this contradiction in her *Frame-and-Compare Evaluation* (FACE) Model. She combines the *Interpretation/Comparison Model* and the SAM by assuming that similarity plays an important role during the judgment process twice.

Summing up, the *Interpretation/Comparison Model* makes many similar assumptions to the SAM and especially to the *Inclusion/Exclusion Model* about when to expect assimilation or contrast effects. Yet, the models offer different underlying processes that lead to assimilation and contrast effects and they also differ regarding the role of similarity of target and standard.

[18] Stapel and Koomen (2001a, p.135) call the *Inclusion/Exclusion Model* an "appropriateness-based model" and say that this class of models does "not concern us here".

The Flexible Correction Model (Wegener & Petty, 1997). This model is focusing on correction processes that people might use when making a judgment. The authors follow the other models presented here inasmuch that they belief that context effects might bias human judgments and lead either to assimilation or contrast effects. However, they contradict the assumption that assimilation effects are the default bias. They rather assume that context effects might cause both, assimilation or contrast effects, in uncorrected judgments. This assumption is, of cause, shared by the SAM that even assumes that the same context or standard can lead to assimilation and contrast effects. With the *Flexible Correction Model*, Wegener and Petty do not try to explain these 'default' or 'uncorrected' biases. Instead, they focus on an additional process that might influence the final judgment: The process of correction.

The *Flexible Correction Model* is not the only theoretical approach that deals with correction processes in judgment. The *Set/Reset Model* (Martin & Achee, 1992) is an additional and especially prominent approach to explain assimilation and contrast effects[19] in judgments. However, the *Flexible Correction Model* is a broader approach that seems to include the *Set/Reset Model*. The basic assumption of the model is that judges possess naïve theories about how context effects might influence their judgments. The model is called flexible because these individual theories or beliefs of biasing influences can include any direction and magnitude of that bias. Thus, people may hold theories for assimilative, contrastive or even no influences of a certain context. These theories can be retrieved from memory or be built on the spot while forming a judgment. If judges are motivated and able to identify potential sources of bias, they will use these theories to correct their initial judgments. Corrections are defined as "the result of people consulting their naïve theories of how potentially biasing factors might influence or have already influenced their perception of the target" (Wegener & Petty, 1997, p. 142).

Theory-guided corrections work in a direction opposite to the perceived bias and in a magnitude commensurate with the perceived magnitude of the bias. Although corrections require more cognitive and motivational resources than non-corrections, this does not mean that (only) effortful judgments always include correction processes. Additionally, these naïve individual theories do not necessarily make the judgments of a target more accurate. The theories do not have to correspond with the actual influences of biasing factors and might, therefore, differ regarding size and direction of a bias. If judges are using their theories for correcting,

[19] The *Inclusion/Exclusion Model* (Schwarz & Bless, 1992a) can also be understood as a correction model. Yet, to my opinion, this is only true for part of that model.

this may only lead to an accurate judgment if these theories are correct. If the theories are not correct regarding the size of the biasing influence, this may either lead to under- or overcorrection. Finally, if the theories are not correct regarding the direction of the biasing influence, this may even lead to exacerbation of the initial bias. Thus, when it comes to assimilation and contrast effects, correction processes might diminish, exacerbate or even reverse an initial judgmental bias.

The authors state that these correction processes do not only directly influence the evaluation of a target, but also in a more indirect way by ‘preemptive corrections’. Instead of only changing a formed judgment, people may also try to avoid a bias by changing their way of searching for and interpreting information, or by avoiding the biasing factor altogether.

The authors present an impressive body of empirical evidence supporting their model. First, they could demonstrate that people hold (correct) believes that contexts sometimes have assimilative (e.g., the effect of mood on judgment of the pleasantness of everyday activities) and sometimes contrastive (e.g., seeing a group of very attractive people before judging average-looking people) influences on judgments (Petty & Wegener, 1993, study 1). Thus, people hold, indeed, different naïve theories about biases. Second, the authors could show that people use these theories to correct their judgments when instructed to avoid context influences (Petty & Wegener, 1993, study 2). This could either decrease the original (or uncorrected) bias or even reverse it when judges were overcorrecting. Thus, in a context where assimilation effects where expected, judges showed contrast effects in their judgments when correcting, but in a context where contrast effects were expected they showed assimilation effects (Wegener & Petty, 1995). Additionally, the authors could even demonstrate that people show different correction processes for the same target when presented in different (i.e., assimilative vs. contrastive) contexts (Wegener, Petty, & Dunn, 1998). Finally, Wegener and Petty (1995, study 4) could show that peoples’ ideographic theories of bias predict their corrections. They found evidence that participants’ corrections were associated with both direction and magnitude of the perceived bias.

These findings already proof that correction processes may be an additional source of assimilation and contrast effects in human judgments. Therefore, the *Flexible Correction Model* can be seen as a supplement for the other models described here, such as the SAM. The *Flexible Correction Model* suggests that correction processes may have an influence on any kind of judgments. People correct their judgments whenever they are motivated and able to do so. Since people hold individual beliefs about biases and may differ in their motivation and ability to

identify potential sources of bias, these correction processes may be a source of error-variance when researcher are interested in the ‘uncorrected’ or ‘default’ influence of the context. It seems, therefore, important to detect the variables that moderate peoples motivation and ability to identify potential sources of bias. In the research conducted by Wegener and Petty, they usually directly instruct participants to correct for biasing influences of context variables. Yet, they explicitly state that corrections also appear with more subtle cues, e.g., by merely bringing the context variable to mind (e.g., Schwarz & Clore, 1983), by changing the wording of questions asked (e.g., Strack, Martin, & Schwarz, 1988), or by changing the instructions given (e.g., Martin, 1986; Petty & Wegener, 1993).

To sum up, although these different models presented here appear similar in their assumptions and make similar predictions concerning the question when to expect assimilation and contrast effects, there are some important differences regarding the underlying processes that may cause assimilation and contrast effects. The most important difference is that the SAM focuses on the stage of target knowledge activation, whereas the other models described here focus on the stage of incorporating this activated knowledge into an evaluation of the target (including correction processes). “The main objective of the selective accessibility model is to describe the mechanisms that underlie the activation of target knowledge during a comparison” (Mussweiler, 2003a, p. 483) Therefore, the other models described here do not have to be seen as mutually exclusive, but, on the contrary, as supplementary to the SAM by offering insight into different stages that influence the outcome of comparative judgments: “Assimilation and contrast effects may reflect the operation of a number of different processes, which have often been conceptualized in independent theories” (Bless et al., 2003, p. 181)

1.5 Goals

In the following part, I want to explain the main goals of my dissertation project. These goals are, on the one hand, closely related to the questions raised at the beginning of this chapter. They are, on the other hand, related to the empirical part of my work that will be presented in the following two chapters. The first goal will be addressed in the second chapter; the second goal will be addressed in the third chapter.

1.5.1 Goal 1

The first goal addresses the question whether prior judgments will have a systematic influence on subsequent ones in a sequential judgment situation. This question is extended by the question for the direction of this influence: Will prior judgments have an assimilative or a contrastive influence on subsequent ones? Finally, these questions should be answered using complex judgment tasks, an important feature of applied judgment situations. As I have outlined above, although the research on sequential performance judgments clearly gives evidence for a systematic influence on prior judgments on subsequent ones, the direction of this effect (at least in applied settings) shows no clear pattern. There seems to be no answer to the question of when to expect assimilation and when to expect contrast effects. One reason for not having an answer to this question may lie in the fact that most of the research seems to be detached from recent theoretical work or models explaining assimilation and contrast effects. In addition, these models, coming from fundamental research, are rarely tested in complex, applied judgment situations.

The first major goal of the work presented here is, therefore, to take a recent theoretical framework (SAM) and to derive assumptions from this framework about when to expect assimilation and contrast effects in sequential judgments. As a sequential judgment situation, I chose the situation of grading exams, a complex judgment situation known and relevant to most people from either one or both sides. I chose the grading of exams for three reasons: First, there seems to be a surprisingly small body of research on how judgments in this field are formed; second, judgments in this situation are complex, relevant and important for any person during one or more times in his educational life; third, exams are usually graded in a sequence, thus, making comparisons between the different targets to be judged most likely.

1.5.2 Goal 2

Apart from testing if there are systematic effects of prior judgments on subsequent ones, I am also interested in the underlying processes that may explain these effects. From my own experience, I derived the assumption that comparisons of different performances may play an important role here, an assumption that is shared with the SAM. The SAM offers a very good insight into the processes and variables underlying assimilation and contrast effects in social judgments. However, the important and central assumption – the idea that judges always follow, test and in most cases confirm (either) a (similarity- or dissimilarity-) hypothesis when comparing a target to be judged with a given standard – has not yet been tested directly.

As I have outlined above there has been research showing that specific information consistent with the assumingly tested hypothesis is more accessible *after* judging the target. Yet, making the judgment could also have activated this information. There is only one study, using think aloud protocols, that shows that information consistent with the hypothesis being tested is being activated (Mussweiler & Strack, 1999). Apart from this, there are no studies examining the postulated process of hypothesis testing by focusing on the kind of information judges are looking at. It seems especially interesting to see whether judges, indeed, follow a certain test-strategy (e.g., a positive test strategy, Klayman & Ha, 1987) as proposed by the SAM. A direct test of this process proposed by the SAM is the second goal of my dissertation project.

As a judgment situation, I chose the situation of oral exams. In this situation teachers have to find out how good the knowledge of a student in a certain area is. To do this, teachers have to look for information by asking the students questions. Therefore, by examining the questions teachers ask during an oral exam, I expect to find evidence for a certain (e.g., positive or confirmatory) test strategy, provided that teachers really follow and test a hypothesis when examining students. Additionally, as explained above, oral exams are also usually graded in a sequence, thus, making comparisons between the different targets to be judged most likely.

These two goals, outlined here, will be addressed by two lines of studies in the following two chapters, including the precise hypotheses these studies are built on. The last chapter will discuss the findings reported in these chapters and give implications following from these results.

Chapter 2
-
Assimilation and contrast effects when grading exams

2.1 Experiment 1: Judgment of written protocols

2.1.1 Overview

The goal of this first line of studies was to investigate whether and in what direction prior judgments would systematically influence the following judgments in a systematic way in a sequential and complex judgment task, an important feature of applied judgment situations. The idea was that prior judgments should function as a comparison standard with which the following targets to be judged would be compared. This comparison should be a possible cause for a judgmental bias in the performance judgments of the target to be judged either in an assimilative or a contrastive way. Unlike other studies concerned with this question, my work was build on a recent theoretical approach – the *Selective-Accessibility Model* (SAM, Mussweiler, 2003a). By combining findings from applied studies with theories based on fundamental research, it may be possible to close the gap between these two fields. The questions relevant to the applied field – namely when to expect assimilation and contrast effects in evaluative judgments – may be answered by fundamental theories, such as the SAM. In addition, it seems to be fruitful to test the assumption of the SAM in more complex judgment situations, i.e., outside the domain of social comparisons and judgmental anchoring.

The SAM specifies that in the case of a given target, assimilation as well as contrast effects may occur in the judgments of the target. In three experiments this idea was tested for the situation of grading exams as a rather complex judgment situation. In this domain teachers usually have to judge more than one pupil or exam in a sequence. It was expected that in this case the first pupil would function as a comparison standard for the following pupil(s) (target[s]). Following the assumptions of the SAM, it was expected that the comparison of two pupils would lead to assimilation effects when teachers follow a similarity hypothesis. In contrast to that,

the same situation would lead to contrast effects when teachers follow a dissimilarity hypothesis.

In the paradigm used in the first three studies, participants were put in the role of a teacher. They had to grade the performance of two pupils in succession in a test in general knowledge. They received written protocols of two pupils, had to read through, and to grade these in the end.

In the first study two variables were manipulated to test the assumptions of the SAM mentioned above. On the one hand, the quality of the standard was manipulated. The exam of the first pupil showed either a good performance (high standard) or a bad performance (low standard). On the other hand, the hypothesis guiding the comparison process was manipulated. Participants were either primed to focus on similarities or on dissimilarities. In accordance with the SAM, an interaction pattern was expected for the evaluative judgments (of the exams) of the second pupil (target). Following a high standard, it was expected that participants would judge the target exams to be better when focusing on similarities (assimilation) than when focusing on dissimilarities (contrast). Following a low standard, participants were expected to judge the target exams to be better when focusing on dissimilarities (contrast) than when focusing on similarities (assimilation).

2.1.2 Method

Participants

We recruited 64 persons at the University of Heidelberg as participants for this experiment, for which they received € 5 in return. Two participants[20] did not work on the task properly and failed to fill out the questionnaire. Three participants reported that they already knew the priming procedure we used from previous studies. The data from these participants was dropped from the analyses. Finally another four participants were students of at least one of the academic fields from which the questions in the written protocols were drawn. Since it was expected that the knowledge of these participants would influence their judgments, the data from these participants was also excluded from the analyses. Summing up, I analyzed the data of 55 (18 female and 36 male) participants. Participants were between 20 and 49 years old (M = 26). 45 of the participants were students in their 1st to 18th semester (M = 6.4). Three participants did not give any demographic information.

[20] Actually, it was the same person participating twice in the same study. Thanks at this point to our research assistants Geoffrey and Sarah who proved to be skilful detectives.

Design

The study followed a 2 (Standard: High vs. low) x 2 (Focus: On similarities vs. on dissimilarities) between participants design. The standard was manipulated by changing the number of correct and incorrect answers in the written protocols. The focus of the participants was manipulated using a procedural priming procedure preceding the grading task. Dependent variables were the grades given to the two pupils, a comparative judgment of the performance and a proposition whom of the two pupils to give an award to in the end.

Materials and Pre-tests

Priming Procedure. A priming procedure was used to get participants into a certain mindset where they would either focus on similarities or dissimilarities in a comparison task. This procedure was adapted from Mussweiler (2001b) and consisted of sketches from two scenes that were taken from Markman and Gentner (1996). In both scenes, a person (in the one a man; in the other a woman) is setting the table for a Christmas celebration (see Appendix 2.A for the pictures). Participants had to compare these two pictures.

Exam Protocols. Participants received written protocols of two pupils. Each one consisted of eight questions and answers. They were seemingly taken from a test in general knowledge. The domains of the questions were chemistry, geography, mathematics and physics. There were two questions and answers for each domain. To manipulate the comparison standard, two different versions of the two protocols were used. In the one version the protocol of the first pupil (standard) showed a good performance (high standard), in the other version the protocols of the first pupil showed a poor performance (low standard). The performance of the pupils was manipulated by changing the number of correct and incorrect answers in the protocols. The well performing standard answered seven of the eight questions correctly, the poorly performing standard answered one of the eight questions correctly. The protocol from the second pupil (target) was the same in both versions and there were four correct and incorrect answers. The two pupils were introduced as two female pupils named Christina and Katrin, 18 and 19 years old, respectively (see Appendix 2.B for the protocols).

Pre-tests. The difficulty of the questions used in the protocols was pre-tested. In the pre-test, 23 participants judged the difficulty of 40 general knowledge questions. Participants were students of the University of Heidelberg and were,

therefore, taken from the same population as the participants of the main study. Only fairly easy questions were chosen for the protocols in the main study. This was done to make sure that participants would be able to distinguish correct and incorrect answers. The difficulty of the questions chosen varied between 1.00 and 3.31 (M = 1.92, SD = 0.62) on a scale from 1 (*very easy*) to 7 (*very difficult*). There were two questions from each knowledge domain in each of the two protocols and the questions for each protocol were matched regarding their level of difficulty (M_{Prot1} = 1.85 vs. M_{Prot2} = 1.99; $t(22) = -1.34$, $p >$. 19).

The answers were written paragraphs between 109 and 192 words of length (M = 144, SD = 20.85). The answers were written in a way that the ones used for the protocols of the standard (first protocol) were obviously correct or incorrect. This was done to ensure that the manipulation of the quality of the standard would work properly. Contrary to that, the answers for the protocol of the target (second protocol) were written in a way that did not look quite so obviously correct or incorrect. This was done to give participants some room for interpretation, which makes a bias in the performance judgments more likely to appear. This was tested in a second pre-test where participants had to judge whether the answers to the questions were correct or not. In the second pre-test, 10 participants judged each version of the two protocols, respectively. Participants were students at the University of Heidelberg and, therefore, taken from the same population as participants in the main study. The difference between the percentage of participants judging an answer as correct and the percentage of participants judging an answer as incorrect was calculated. The resulting score could range from 0 (not easily identifiable as correct or incorrect) to 100 (easily identifiable as correct or incorrect). Pooling the ratings of both versions, participants identified answers more easily as correct or incorrect for the standard (M = 87.50) than for the target protocols (M = 73.75), $t(19) = 2.42$, $p < .05$.

Participants also indicated the confidence of their judgments on a scale from 1 (*not confident*) to 7 (*very confident*). Pooling the ratings of both versions, participants were more confident about their ratings for the standard (M = 5.94) than for the target protocols (M = 4.90), $t(19) = 5.91$, $p < .001$.

Procedure

When participants arrived at the laboratory, they first filled out an informed consent form. They were then given the two pictures of the priming task. In the instructions they were asked to compare these two pictures. Half of the participants were instructed to find similarities between the two pictures (focus on similarities). The other half of the participants was instructed to find differences between the two

pictures (focus on dissimilarities). All participants were told that they should take a couple of minutes for this task and that they should try to find all the similarities / differences between the two pictures. Participants then had to write down all the similarities / differences they could find. This task was taking about 10 minutes on average. After that they were told about a second task, the judgment of the examination protocols. They were told that they would have to read through protocols of two pupils. Both pupils had allegedly been proposed for an award. Participants were instructed to judge the performance of the two pupils by grading the protocols on a scale from 1.0 (*very good*) to 5.0 (*very bad*)[21]. Half of the participants were given the version with the well performing first pupil (high standard), the other half were given the version with the poorly performing first pupil (low standard).

In addition to the grades, participants judged the performance of the target compared to the standard on a scale from 1 (*a lot worse*) to 7 (*a lot better*). They also indicated, which of the two pupils they would propose for the award. In the end they also had to indicate their gender, age, what they were studying (if they were students), and in which semester they were. Finally, they were fully debriefed and thanked for their participation by the research assistants running the study.

2.1.3 Results

In a preliminary analysis, I found the age of the participants to be correlated with the dependent variables (the grades of the two pupils). Older people tended to give inferior grades[22] for the first (r = .29; p < .05) as well as for the second pupil (r = .35; p < .05). This variable was, therefore, used as a covariate in the following analyses.

As a manipulation check, I first looked at which one of the pupils participants proposed for the award. As expected, most participants (84 %) proposed the first pupil for the award in the high standard condition. In the low standard condition, most participants (81 %) proposed the second pupil for the award.

As a second manipulation check, I performed a standard x focus analysis of variance (ANOVA) on the performance judgments (grades) of the first pupil. As expected, the high standard (M = 1.52) was given better grades than the low standard (M = 3.63), $F(1,48) = 93.86$, p < .001, η^2= .66. Unexpectedly, there was also

[21] This is the normal grading scale used at most German schools and universities and was therefore highly familiar for participants.
[22] The grading system common in German schools was used. The grades range from 1 to 5 and lower numbers correspond with better grades.

a significant interaction, F (1,48) = 6.94, p < .05, η^2 = .13. Participants judged the performance of the high standard to be better when focusing on similarities (M = 1.39) than when focusing on dissimilarities (M = 1.65). For the low standard the opposite pattern emerged: The performance was judged better when focusing on dissimilarities (M = 3.28) than when focusing on similarities (M = 4.00). This interaction is displayed in figure 2.1.

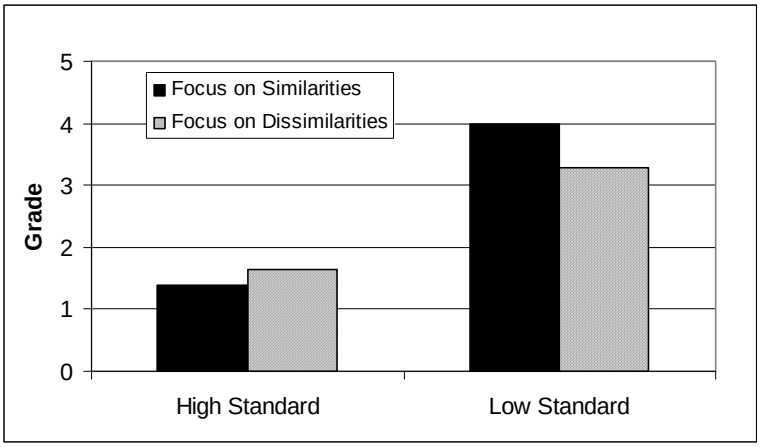


Figure 2.1. Performance judgments of the first pupil (standard)

To test the main hypothesis, I performed a standard x focus analysis of variance (ANOVA) on the performance judgments (grades) of the second pupil. There were no significant effects, but there was a tendency for an assimilation effect. Participants judged the performance of the target to be better after a high than after a low standard F(1,48) = 2.44, p < .13, η^2 = .05. More importantly, the results also show the expected interaction pattern, F(1,48) = 2.58, p < .12, η^2 = .05 (see figure 2.2). After a high standard participants judged the performance of the target to be better when focusing on similarities (M = 2.15) than when focusing on dissimilarities (M = 2.42). For the low standard the opposite pattern emerged: The performance was judged to be better when focusing on dissimilarities (M = 2.58) than when focusing on similarities (M = 2.96). There was no independent effect for the focus (F < 1). An analysis of the simple effects revealed a significant difference in the judgments after high compared to low standard when focusing on similarities (F(1,48) = 5.02, p < .05, η^2 = .10), but no significant difference when focusing on dissimilarities (F < 1).

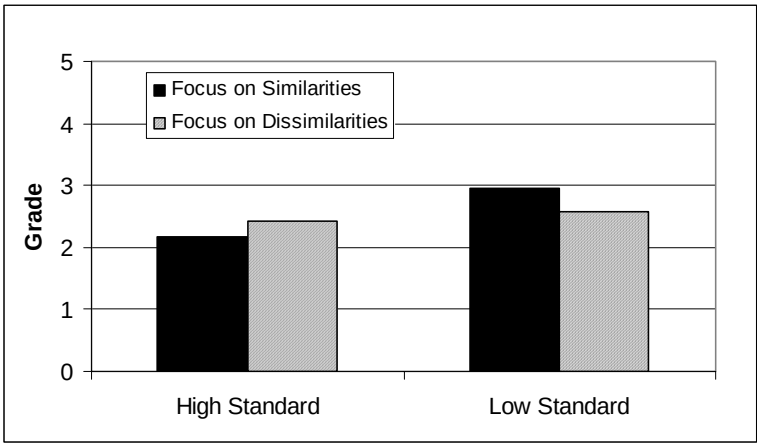


Figure 2.2. Performance judgments of the second pupil (target)

As a second test of the main hypothesis, I re-ran the analysis with the relative ratings as dependent variable. An interaction was expected, but only a trivial main effect emerged ($F(1,48) = 29.14$, $p < .001$, $\eta^2 = .38$), showing that the performance of the second pupil was judged worse[23] than the first pupil in the high standard condition ($M = 2.85$); in the low standard condition the performance of the second pupil was judged better than the first pupil ($M = 4.74$). This main effect can be interpreted as an additional manipulation check.

2.1.4 Discussion

This study was run to test whether and in what direction prior judgments would systematically influence following judgments in a sequential and complex judgment situation. The study was based on a theoretical framework (the SAM) and it was additionally tested whether the assumptions of the SAM (Mussweiler, 2003a) also apply for the complex task of making performance judgments in an exam situation. Following the assumptions of the SAM, it was expected that participants would show assimilation effects in their judgments of two pupils when following a similarity hypothesis, but would show contrast effects when following a dissimilarity hypothesis. To test this, participants were either primed to focus on similarities or dissimilarities. They then had to judge the performance of two pupils. The first one (standard) was either performing very well (high standard) or very poorly (low standard). This was expected to influence the judgment of the second pupil (target). It was expected that the target would be judged similar to the standard when participants were focusing

[23] On a scale from 1 (*a lot worse*) to 7 (*much better*)

on similarities (assimilation) and different to the standard when participants were focusing on dissimilarities (contrast).

Overall, this pattern was found in this first study. After a high standard the target was judged better when participants focused on similarities than when they focused on dissimilarities. After a low standard the opposite pattern appeared: The target was judged worse when participants focused on similarities than when they focused on dissimilarities. This pattern shows that a focus on similarities leads to assimilation effects and a focus on dissimilarities leads to contrast effects in the judgments of the target, as predicted by the SAM.

This interaction failed to reach statistical significance. One reason for this may be found in the material used in this study. Maybe the written protocols did not leave enough room for interpretation of the target performance. Confirmation of a hypothesis is more likely to occur when different information may be retrieved (e.g., activated from memory) prior to the judgment or when the given information is ambiguous. In the latter case, information may be encoded in terms of what is being expected (Trope & Liberman, 1996). In the paradigm used, all participants retrieved exactly the same information. Therefore, no bias could be expected at the stage of retrieval. However, a bias was still expected to occur at the stage of interpretation. One reason why the bias in the judgments was not as strong as expected might be that the room for interpretation was not big enough. This might be possible, although the materials were chosen on the basis of a pre-test trying to ensure that the answers to judge were highly ambiguous. The room for interpretation or the ambiguity of the performance to be judged might still be a crucial variable that determines the size of the bias. Therefore, in the following study the paradigm was changed in a way to increase the ambiguity of the performance of the target.

A second explanation why the biasing effects were not as strong as expected could be that the priming manipulation did not have the expected effect. The interaction pattern (unexpectedly) found in the judgments of the standard makes this rather unlikely. The same pattern was found for the judgments of the standard as for the judgments of the target. The high standard was judged better when participants focused on similarities than when they focused on dissimilarities. For the low standard the opposite pattern appeared: The low standard was judged worse when participants focused on similarities than when they focused on dissimilarities. The SAM can explain this finding if one assumes that participants were building a standard of comparison while reading through the protocols. For there was no standard given for the judgment of the first pupil, participants were maybe using the performance at the beginning of the protocols as a standard to compare the rest of

the performance with. A second idea would be that participants used their own abilities as a standard (c.f. Biernat, Manis, & Kobrynowicz, 1997). Participants may also have used some abstract idea of how an average pupil would or should perform as a comparison standard. However, in this case, only a main effect would have been expected. So, it seems plausible that participants first tried to get some idea about the abilities of a pupil they had to grade. After they had developed an idea or hypothesis, participants tested this hypothesis following a certain (positive) test-strategy as suggested by the SAM. An important additional point of this finding is that it shows that the priming manipulation had the expected effect. Still, it could be possible that these effects decreased over time. This would explain the smaller effects in the judgments of the target (second pupil).

To conclude, the first study showed that when exam performances of two pupils are judged in a sequence, the first performance judgment influences the following one. Following the assumptions of the SAM, it seems that the first pupil functions as a standard of comparison for the following one (target). Depending on the hypothesis participants follow during the comparison of target and standard, this leads either to assimilation or contrast effects in the judgments. The size of this bias depends on information activated during the comparison process and the ambiguity of the information at the stage of interpretation. To replicate the findings of study one, in the following study the information given about the target were made more ambiguous.

2.2 Experiment 2: Judgment of audio recorded protocols

2.2.1 Overview

When evaluative performance judgments are made in a sequence, prior judgments may influence the following ones either in an assimilative or in a contrastive way. Whether assimilation or contrast effects occur depends on the hypothesis people follow when making judgments according to the SAM (Mussweiler, 2003a). In the first experiment, this assumption was tested for the judgments of exams. It was found that the performance of a first pupil served as a comparison standard for the following one. When participants were focusing on similarities, assimilation effects occurred in the performance judgments of the target. When participants were focusing on dissimilarities, contrast effects occurred in the performance judgments of the target. Yet, these effects were too weak to reach statistical significance. A bias in the judgments is more likely to occur if the performance is ambiguous and can be interpreted in different ways. Therefore, in this second study the written protocols of the first study were “translated” into audio-recorded protocols. One main difference between reading through written protocols and listening to audio-recorded protocols is the lack of control in the latter. Whereas participants can control their reading speed, stop at times to think about the answers, and even reread some answers in the case of written protocols, they do not have this opportunity while listening to audio recordings. This should make the audio recordings more ambiguous than the written protocols. So, it was expected to replicate the effects found in study one and these effects were expected to be stronger in study two. The task of participants was the same as in study one. They were in the role of a teacher and had to grade the performance of two pupils taking an exam testing general knowledge.

2.2.2 Method

Participants

We recruited 63 students at the University of Heidelberg as participants for this experiment, for which they received € 4 in return. Five participants did either not work on the task as instructed or were given the material in a wrong order. The data from

these participants was dropped from the analyses. Another four participants were students of at least one of the academic fields from which the questions in the exams were drawn. Since it was expected that the knowledge of these students would influence their judgments, these students were also excluded from the analyses. Summing up, I analyzed the data of 54 (34 female and 20 male) participants. Participants were between 20 and 29 years old (*M* = 23). All of the participants were students, studying in their 1st to 14th semester (*M* = 2.7).

Design

The study followed a 2 (Standard: high vs. low) x 2 (Focus: On similarities vs. on dissimilarities) between participants design. The standard was manipulated by changing the number of correct and incorrect answers in the audio-recorded protocols. The focus of the participants was manipulated using a procedural priming task preceding the grading task. Dependent variables were the grades given to the two pupils, a comparative judgment of the performance, a proposition whom of the two pupils to give the award to, and a general judgment of the ability of all pupils in Germany.

Materials

Priming Procedure. The priming procedure was the same as in study one. Two pictures had to be compared. The task for participants was either to focus on similarities or on differences between the two pictures.

Exam Protocols. The written protocols of study one were used to create audio recordings of the alleged exams. In these audio recordings one could hear two people, the teacher asking questions and the respective pupil giving answers. The quality of speech was the same for both the target and the (two) standard(s). The two examinees were introduced as two pupils named David and Christian. As in study one, there were two different versions of audio recordings to manipulate the performance of the standard. In the one version, the first pupil was performing well (high standard), in the other version the first pupil was performing poorly (low standard). The performance of the pupils was manipulated by changing the number of correct and incorrect answers. The well performing pupil answered seven of the eight questions correctly; the poor performing pupil answered one of the eight questions correctly. The performance of the second pupil (target) was the same in both versions. The second pupil answered four of the eight questions correctly. In both versions, the same actor was used for the voice of the first pupil (standard), but

a different actor was used for the voice of the second pupil (target). The actor for the voice of the teacher was the same for all recordings.

Procedure

When participants arrived at the laboratory, they first filled out an informed consent form. They were then given the two pictures of the priming task that was used in study one (see 2.1.2). Half of the participants were instructed to find similarities between the two pictures. The other half of the participants was instructed to find differences between the two pictures. Following this task, participants were told about a second task, the judgment of the audio-recorded protocols. They were told that they would listen to audio-recorded protocols of two pupils. Both pupils have allegedly been proposed for an award. Participants were instructed to judge the performance of the two pupils by grading the audio recordings on a scale from 1.0 (*very good*) to 5.0 (*very bad*)[24]. Half of the participants were given the version with the well performing first pupil (high standard), the other half were given the version with the poorly performing first pupil (low standard).

Additional to the grades, participants judged the performance of the target compared to the standard on a scale from 1 (*a lot worse*) to 7 (*a lot better*). They also indicated, which of the two pupils they would propose for the award. Finally, participants had to judge the ability of all students in Germany concerning general knowledge. In the end, they also had to indicate their gender, age, what they were studying (if they were students), and in which semester they were. Finally, they were fully debriefed and thanked for their participation by the research assistants running the study.

2.2.3 Results

In a preliminary analysis, I found gender of the participants to be correlated with the dependent variables (the grades of the two pupils). Male participants tended to give inferior grades[25] (M = 2.97) compared to female participants (M = 2.27) for the first ($t(52) = -1.91$; $p < .06$) as well as for the second pupil (M_{male} = 2.44 vs. M_{female} = 2.02; $t(52) = -1.51$, $p < .14$). This variable was, therefore, used as a covariate in the following analyses.

[24] This is the normal grading scale used at most German schools and universities and was therefore highly familiar for participants.
[25] The German school grading system was used. The grades range from 1 to 5 and lower numbers correspond with better grades.

As a manipulation check, I first investigated who of the two pupils participants proposed for the award. As expected, most participants (85 %) proposed the first pupil for the award in the high standard condition. Unexpectedly, in the condition with the low standard only 44 % of the participants proposed the second pupil for the award.

I also performed a standard x focus analysis of variance (ANOVA) on the performance judgments (grades) of the first pupil as a second manipulation check for the manipulation of the standard. As expected, the high standard (*M* = 1.58) was given better grades than the low standard (*M* = 3.48), $F(1,49) = 56.70$, $p < .001$, $\eta^2 = .54$ (see figure 2.3). No other significant effects were found (all *F*s < 1). Taken together, both analyses suggest that the manipulation of the standard worked in the expected direction.

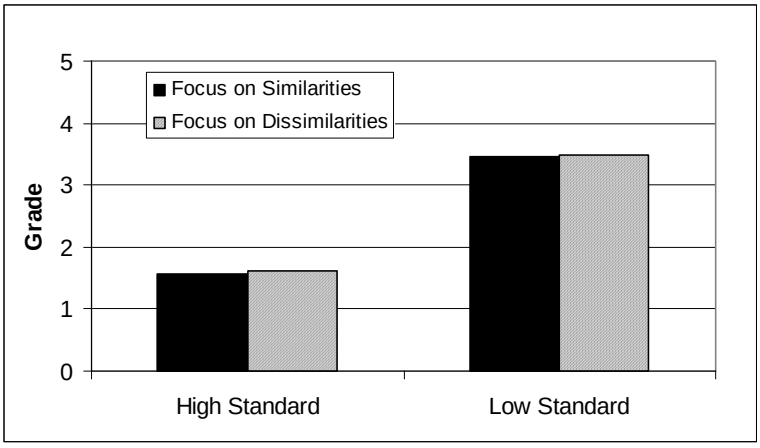


Figure 2.3. Performance judgments of the first pupil (standard)

To test the main hypothesis, I performed a standard x focus analysis of variance (ANOVA) on the performance judgments (grades) of the second pupil (target). An assimilation effect in the judgments emerged. Participants judged the performance of the target to be better following a high (*M* = 1.89) than following a low standard (*M* = 2.45), $F(1,49) = 4.18$, $p < .05$, $\eta^2 = .08$. There were no other significant effects (all *F*s < 1). Especially, the expected interaction did not appear.

As a second test of the main hypothesis, I re-ran the analysis with the relative ratings as dependent variable. An interaction was expected, but only a trivial main effect emerged ($F(1,49) = 11.20$, $p < .01$, $\eta^2 = .19$), showing that the performance of the second pupil was judged worse[26] than the first pupil in the high standard

[26] On scale from 1 (*a lot worse*) to 7 (*much better*)

condition (*M* = 3.81); in the low standard condition, the performance of the second pupil was judged better than the first pupil (*M* = 5.01). This main effect can be seen as an additional manipulation check.

A third dependent variable was the judgment of the ability of all pupils in Germany. Overall, participants judged the ability of the German pupils concerning general knowledge to be rather poor[27] (*M* = 3.35). In comparison with the mean judgments of the first (*M* = 2.53) and the second pupil (*M* = 2.17), the judgments of German pupils in general was worse, $t(53) = 5.62$, $p < .001$.

To test whether the independent variables also influenced the judgments of all pupils in Germany, the same analysis as before was run with the accordant judgments as dependent variable. It was expected that the biases in the judgments of the two pupils would carry over to the judgment of all pupils in Germany. I found a tendency of a contrast effect, i.e., for participants to give worse grades after a high standard (*M* = 3.49) than after a low standard (*M* = 3.21), $F(1,49) = 2.11$, $p < .16$, $\eta^2 = .04$. There was also a tendency for participants to give worse grades when focusing on similarities (*M* = 3.50) than when focusing on dissimilarities (*M* = 3.18), $F(1,49) = 2.51$, $p < .13$, $\eta^2 = .05$. The interaction was not significant ($F < 1$). Taken together, there was a tendency for contrast effects in the ability judgments of all German pupils.

2.2.4 Discussion

As the first study, this second study was run to test whether in a sequential judgment situation prior judgments would systematically influence following judgments. The study was based on a theoretical framework (the SAM) and it was additionally tested whether the assumptions of the SAM (Mussweiler, 2003a) also apply for the complex setting of making performance judgments in an exam situation. Following the SAM, it was expected that participants would show assimilation effects in their judgments when following a similarity hypothesis, but would show contrast effects when following a dissimilarity hypothesis. To test this, participants were either primed to focus on similarities or dissimilarities. They then had to judge the performance of two pupils. The first one (standard) was either performing very well or very poorly. This was expected to influence the judgments of the second pupil

[27] The German school grading system was used. The grades range from 1 to 5 and lower numbers correspond with better grades.

(target). It was further expected that the target would be judged similar to the standard when participants were focusing on similarities (assimilation) and different to the standard when participants were focusing on dissimilarities (contrast). In this second study, participants were judging the performance of the pupils after listening to audio recordings of two exams. It was expected that this would increase the amount of ambiguity of the performance compared to study one where written protocols were used instead. A higher amount of ambiguity should make a judgmental bias more likely to appear.

Although a systematic influence of prior judgments on subsequent ones was found, the expected interaction pattern did not occur in this study. Instead, an assimilation effect in the performance judgments of the second pupil (target) appeared. After a well performing first pupil (high standard), participants judged the performance of the target to be worse than after a poorly performing first pupil (low standard). One possible explanation for this result might be that the priming task failed to influence the focus of comparison as expected, although it worked properly in the first study. This might be due to the different material used in this study. Compared to the first study, participants spent a longer time listening to the audio-recorded exams than reading through the written protocols in study one[28]. So, the effects of the priming manipulation might have weakened over time. For I did not find the same interaction pattern as in the first study in the judgments of the standard, this seems to be a plausible explanation. The time lag between the manipulation of the focus and the judgment of the standard was too long for the manipulation to have some effects. Therefore, the focus of comparison participants followed was determined by some alternative source and not by the priming procedure used. If one assumes that both the standard and the target came from the same peer group (both pupils) and that they were both proposed for winning an award, this might have led participants to see both of them to be rather similar. If this was the case, participants would be following and testing a similarity hypothesis in the following judgments in all of the conditions. According to the SAM, this would result in assimilation effects. This is what I found in the performance judgments of the target.

Participants also had to judge the ability of all pupils in Germany concerning general knowledge. In these judgments contrast effects (compared to the standard) appeared. There was a tendency for participants to judge the ability of all pupils to be worse after a high than after a low standard. There was also a tendency for

[28] There was even an additional time lag between the exams of the two pupils, because the recordings of the two exams were on two sides of the same tape. Therefore the tapes had to be fast-forwarded between the two exams.

participants to give worse ability judgments when focusing on similarities than when focusing on dissimilarities, but I do not have a conclusive for this latter finding. It seems that the priming manipulation had some unexpected side effects. This might also be due to the long time lag between the initial priming task and this final judgment. As for the contrast effects in the judgments of all pupils in Germany, these are not as unexpected if one considers the preceding judgments of standard and target. Again, if one assumes that the effects of the priming manipulation had disappeared, participants would still judge target and standard in a quick holistic judgment to be either similar or different. Following the same line of arguments used above, maybe the two pupils judged before were seen as rather exceptional pupils because both of them were proposed to win an award. Therefore, they were both belonging to a category of exceptional pupils that differed from the category of all ‘normal’ pupils in Germany. In this case, the standard and the target would be seen as being rather different and this would lead to a focus on dissimilarities and contrast effects in the ability judgments.

Taken together, the findings of the second study can be explained if one assumes that the two pupils to be judged were seen as belonging to the same category of ‘exceptional pupils’ that is different to the category of ‘normal pupils’ in Germany. To conclude, study two again gives evidence that when two exams are graded in a sequence, the first judgment influences the following one. Contrary to the first study, this time only assimilation effects were found to appear in the judgments of the second exam. Even so contrast as well as assimilation effects were expected, these results do not necessarily contradict the assumptions of the SAM. It seems likely that the priming manipulation was too weak to influence the hypothesis or focus of participants in the expected way. The SAM assumes that the comparison of standard and target is always guided by a hypothesis. However, the model does not make any explicit statements about when this hypothesis assumes target and standard to be similar or to be dissimilar. In the paradigm used in study two, the information that both pupils were proposed for an award might have been seen as an important cue for similarity, inducing participants to follow and test a similarity hypothesis. This explains why only assimilation effects occurred in the judgments of the second exam, and contrast effects in the judgments of all ‘normal’ pupils in Germany. A reason for the priming procedure not to work as well in the second as in the first study might have been the long temporal distance between the priming task

and the judgment of the two pupils. Therefore, in a following third study I try to increase the impact of the priming procedure on the judgments of the target by placing it closer to the judgments of the target in the experimental setting.

2.3 Experiment 3: Judgment of audio recorded protocols (part II)

2.3.1 Overview

In a sequential judgment situation, prior judgments may systematically influence the following ones either in an assimilative or in a contrastive way. Whether assimilation or contrast effects occur depends on the hypothesis people follow when making judgments according to the SAM (Mussweiler, 2003a). In the first two experiments, this assumption was tested for the judgment of exams. In experiment one, it was found that the performance of a first pupil served as a comparison standard for the following one (target). When participants were focusing on similarities, assimilation effects occurred in the performance judgments of the target; when participants were focusing on dissimilarities, contrast effects occurred in the performance judgments of the target. Yet, these effects were not as strong as expected. Therefore, in a second study the performances to be judged were presented as audio recordings of the exams rather than as written protocols. It was expected that this would make the performance to be judged more ambiguous. Again, it seems that the first pupil was used as a comparison standard for the following one, but this time resulting in assimilation effects in the judgments of the second pupil (target)[29]. These results suggest that the priming procedure used to manipulate the focus of comparison did not have the expected effects. Instead, participants seemed to view the target and standard as being similar to each other in all conditions. The priming procedure might have been too weak in the second study because of the long time lag between the manipulation and the judgment of the target. To strengthen the influence of the priming procedure, this time lag was reduced in the following study. The priming procedure was relocated in the course of the experiment. It was now situated between the two exams to be judged, i.e., exactly after grading the first pupil and before starting to listen to the audio recording of the second pupil.

[29] These effects were found for all participants independent of the manipulation of the comparison focus.

2.3.2 Method

Participants

We recruited 67 students at the University of Heidelberg as participants for this experiment, for which they received € 4 in return. Three participants did not work on the task as instructed. The data from these participants was dropped from the analyses. Another two participants were students of at least one of the academic fields from which the questions in the written protocols were drawn. Since it was expected that the knowledge of these students would influence their judgments these students were also excluded from the analyses. Summing up, I analyzed the data of 62 (30 female and 31 male) participants[30]. Participants were between 18 and 31 years old (*M* = 23). All of the participants were students, studying in their 1st to 14th semester (*M* = 4.7).

Design

As in the previous two studies, this third study followed a 2 (Standard: High vs. low) x 2 (Focus: On similarities vs. on dissimilarities) between participants design. The standard was manipulated by changing the number of correct and incorrect answers in the audio-recorded protocols. The focus of the participants was manipulated using a procedural priming task preceding the exam of the second pupil. Dependent variables were the grades given to the two pupils, a comparative judgment of the performance, a proposition whom of the two pupils to give an award to, and a general judgment of the ability of all pupils in Germany.

Materials

Priming Procedure. The priming procedure was the same as in the first two studies. Two pictures had to be compared. The task for participants was either to focus on similarities or on differences between these two pictures.

Exam Protocols. The same audio recordings as in study two were used (see 2.2.2). As in the previous studies, there were two different versions of audio recordings. In the one version, the first pupil was performing well (high standard), in the other version the first pupil was performing poorly (low standard).

[30] One participant did not give demographic information.

Procedure

When participants arrived at the laboratory, they first filled out an informed consent form. In contrast to the first two studies, participants directly started with the grading task in this third study. They first had to listen to the performance of the first pupil (standard). Before that, they were told that they would listen to audio-recorded protocols of two pupils. Both pupils had allegedly been proposed for an award. Participants were instructed to judge the performance of the two pupils by grading the audio recordings on a scale from 1.0 (*very good*) to 5.0 (*very poor*)[31]. Half of the participants were given the version with a well performing first pupil (high standard), the other half was given the version with the poorly performing first pupil (low standard).

After grading the first pupil, participants were given the same two pictures of the priming task used in the previous studies (see 2.1.2). They were told that this task was a pre-test for a different study. They should work on this task while the tape was being fast-forwarded to the audio recording of the second exam. Half of the participants were instructed to find similarities between the two pictures. The other half of the participants was instructed to find differences between the two pictures.

After the priming task, participants listened to the audio recording of the second pupil (target) and graded the performance afterwards. Additional to the grades they judged the performance of the target compared to the standard on a scale from 1 (*a lot worse*) to 7 (*a lot better*). They also indicated which of the two pupils they would propose for an award. Finally, they had to judge the ability of all students in Germany concerning general knowledge. In the end, they also had to indicate their gender, age, what they were studying (if they were students), and in which year (semester) they were. They were then fully debriefed and thanked for their participation by the research assistants running the study.

2.3.3 Results

In a preliminary analysis, I found gender of the participants to be correlated with the dependent variables (the grades of the two pupils). Male participants tended to give inferior grades[32] (M = 2.77) compared to female participants (M = 2.30) for the first ($t(59) = -1.49$; $p < .15$) as well as for the second pupil (M_{male} = 2.59 vs. M_{female} =

[31] This is the normal grading scale used at most German schools and universities and was therefore highly familiar for participants.
[32] The grading system common at German schools and universities was used. The grades range from 1 to 5 and lower numbers correspond with better grades.

2.17; $t(59) = -1.69$, $p < .10$). This variable was, therefore, used as a covariate in the following analyses.

As a manipulation check, I first looked at which one of the pupils participants proposed for the award. As expected, most participants (71 %) proposed the first pupil for the award in the high standard condition. In the low standard condition, most participants (70 %) proposed the second pupil for the price.

I also performed a standard x focus analysis of variance (ANOVA) on the performance judgments (grades) of the first pupil as a second manipulation check for the manipulation of the standard. As expected, the high standard ($M = 1.78$) was given better grades than the low standard ($M = 3.33$), $F(1,56) = 34.21$, $p < .001$, $\eta^2 = .38$. No other significant effects were found (all Fs < 1). Taken together, both analyses suggest that the manipulation of the standard worked in the expected direction (see figure 2.4).

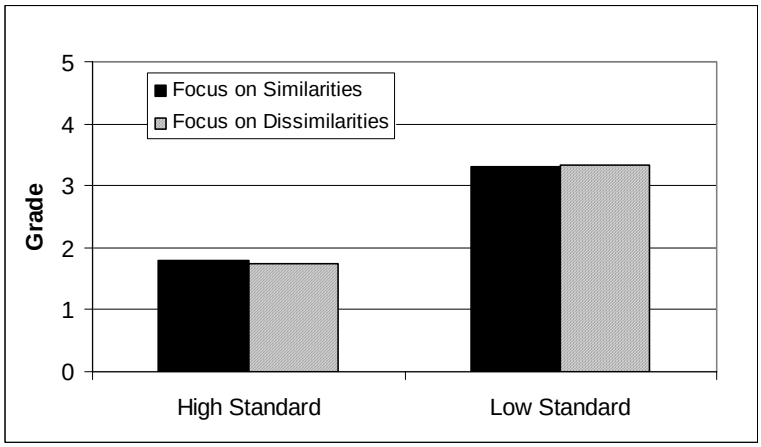


Figure 2.4. Performance judgments of the first pupil (standard)

To test the main hypothesis I performed a standard x focus analysis of variance (ANOVA) on the performance judgments (grades) of the second pupil (target). Unexpectedly, there were no significant effects (all Fs < 1) at all. Especially, the expected interaction did not appear (see table 2.1).

Table 2.1

Performance judgments (grades) of the second pupil (target)

Standard	Focus on	*M*	*SD*	*N*
High standard	Similarities	2.21	1.11	15
	Dissimilarities	2.46	0.96	16
Low standard	Similarities	2.40	0.83	17
	Dissimilarities	2.48	1.19	13

As a second test of the main hypothesis, I re-ran the analysis with the relative ratings as dependent variable. An interaction was expected, but only a trivial main effect emerged ($F(1,56) = 29.73$, $p < .001$, $\eta^2 = .35$), showing that the performance of the second pupil was judged worse[33] than the first pupil in the high standard condition ($M = 3.42$); in the low standard condition the performance of the second pupil was judged better than the first pupil ($M = 5.15$). This main effect can be seen as an additional manipulation check.

A third dependent variable was the judgment of the ability of all pupils in Germany. Overall, participants judged the ability of all pupils in Germany concerning general knowledge to be rather poor[34] ($M = 3.50$). In comparison with the mean judgments of the first ($M = 2.54$) and the second pupil ($M = 2.39$), the judgments of all pupils in Germany were worse, $t(60) = 8.08$, $p < .001$. To test whether the independent variables influenced the judgments of all pupils in Germany, the same analysis as before was run with the accordant judgment as dependent variable. There was a tendency for participants to give worse grades when focusing on similarities ($M = 3.60$) than when focusing on dissimilarities ($M = 3.39$), $F(1,56) = 2.24$, $p < .15$, $\eta^2 = .04$. All other effects were not significant (all Fs < 1).

2.3.4 Discussion

Like the first two studies, this third study was run to test whether and in what direction prior judgments would systematically influence following judgments in a sequential and complex judgment situation. This third study was based on a

[33] On scale from 1 (*a lot worse*) to 7 (*much better*)

[34] The grading system common at most German schools and universities was used. The grades range from 1 to 5 and lower numbers correspond with better grades.

theoretical framework (the SAM) and it was additionally tested whether the assumptions of the SAM (Mussweiler, 2003a) also apply for the complex setting of making performance judgments in an exam situation. According to the SAM, it was expected that participants would show assimilation effects in their judgments when following a similarity hypothesis, but would show contrast effects when following a dissimilarity hypothesis. To test this assumption, participants were either primed to focus on similarities or dissimilarities. They then had to judge the performance of two pupils. The first one (standard) was either performing very well (high standard) or very poorly (low standard). This was expected to influence the judgments of the second pupil (target). It was expected that the target would be judged similar to the standard when participants were focusing on similarities (assimilation) and different when participants were focusing on dissimilarities (contrast). In this study, as in study two, participants were judging the performance of the two pupils after listening to audio recordings of the exams. It was expected that this would enhance the amount of ambiguity of the performance compared to study one where written protocols were used. Additionally, the priming procedure was located exactly before listening to and judging the performance of the second pupil (target). This was done to increase the influence of the priming manipulation compared to study two.

The expected pattern was not found in this study. Instead, no biases in the judgments were found at all. Neither the expected interaction pattern appeared, nor the assimilation effects found in study two could be replicated. Two alternative explanations can be offered for this nil effect. On the one hand, it may be that the first and the second pupil have not been compared in this third study at all. The priming task might not only affect the focus of participants during a comparison task. It is also possible that the priming procedure additionally puts participants into a ‘mindset of comparing’, making comparison effects more likely. Maybe activating this ‘mindset of comparing’ is necessary to find influences of the judgments of the standard on the judgments of the target. On the other hand, it may have been too obvious for participants that the priming procedure was intended to influence their judgments. Participants may, therefore, have tried not to use the first pupil as a standard of comparison. This would have de-biased their judgments. Of course, the SAM assumes that there is no judgment without comparison. Therefore, participants may have used a different standard to compare the target with. However, if these standards were chosen on an individual basis, this would lead to random effects on the aggregate level.

Participants also had to judge the ability of all pupils in Germany concerning general knowledge. There was a tendency for participants to give worse ability

judgments when focusing on similarities than when focusing on dissimilarities. It seems that the priming manipulation had some unexpected side effects. It makes sense that no other effect, especially the contrast effects found in study two, occurred. A bias in this judgment was expected as a carry-over effect from the judgment of the target (second pupil). Since there was no bias in the judgment of the target, no bias should be expected in the ability judgment of all pupils in Germany. Taken together, the findings of these three studies are not easy to be interpreted. Different patterns were found in these slightly different paradigms. To get a different, broader view of the data, I reanalyzed the data of all three experiments together in a ‘mini meta-analysis’. These findings will be presented in the following section.

2.4 Studies 1 to 3: A 'mini meta-analysis'

2.4.1 Overview

The findings of studies 1 to 3 are not easily interpreted. Even if each study on its own does lead to some conclusions, it is a difficult task to come up with some general understanding of the overall pattern found in these studies. To get a broader view, a mini meta-analysis was conducted. The data from the three studies was combined and reanalyzed much in the same way as in the three single studies. It was expected to find a general pattern of results that holds true over the three studies. Since participants in study one received written protocols instead of audio recordings, an additional factor 'protocols' (written vs. audio-recorded) was included in the main analyses. The findings will be discussed in the general discussion.

Participants

In this meta-analysis I analyzed the data of 170 (82 female and 87 male, one missing) participants. Participants were between 18 and 49 years old (M = 24).

2.4.2 Results

In a preliminary analysis, I found gender as well as the age of the participants to be correlated with the dependent variables (the grades of the two pupils). Male participants tended to give inferior grades[35] (M = 2.80) compared to female participants (M = 2.29) for the first ($t(166) = -2.58$; $p < .05$) as well as for the second pupil (M_{male} = 2.51 vs. M_{female} = 2.20), $t(166) = -2.09$, $p < .05$. Older participants tended to give worse grades than younger participants for the first ($r = .12$, $p < .14$) as well as for the second pupil ($r = .19$, $p < .05$). These variables were, therefore, used as covariates in the following analyses.

As a manipulation check, I first looked at the ratings of the participants, which one of the pupils they proposed for the award. As expected, most participants (66 %) proposed the first pupil for the award in the high standard condition. In the condition with the low standard most participants (78 %) proposed the second pupil for the award.

[35] The German school grading system was used. The grades range from 1 to 5 and lower numbers correspond with better grades.

I also performed a standard x focus x protocols analysis of variance (ANOVA) on the performance judgments (grades) of the first pupil as a second manipulation check for the manipulation of the standard. As expected, the high standard (*M* = 1.63) was given better grades than the low standard (*M* = 3.47), *F*(1,158) = 152.04, *p* < .001, η^2= .49. The overall pattern of results was also the same as in study one, suggesting an interaction effect. Participants judged the performance of the high standard to be better when focusing on similarities (*M* = 1.59) than when focusing on dissimilarities (*M* = 1.68). For the low standard the opposite pattern emerged: The performance was judged better when focusing on dissimilarities (*M* = 3.37) than when focusing on similarities (*M* = 3.57), *F* (1,158) = 3.01, *p* < .09, η^2 = .02. However, this effect was qualified by a three-way interaction (*F*(1,158) = 3.40, *p* < .07, η^2 = .02), meaning that this interaction only showed in study one where written protocols were used (see figure 2.5). No other significant effects were found (all *F*s < 1.3). Taken together, both analyses suggest that the manipulation of the standard worked in the expected direction.

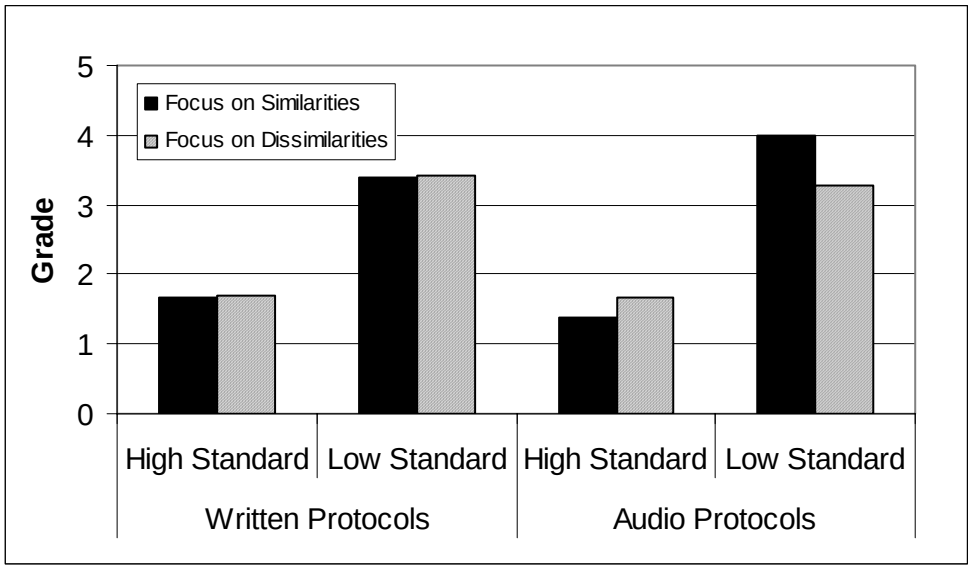


Figure 2.5. Performance judgments of the first pupil (standard)

To test the main hypothesis, I performed a standard x focus x protocols analysis of variance (ANOVA) on the performance judgments (grades) of the second pupil. Overall, an assimilation effect in the judgments emerged (see figure 2.6). Participants judged the performance of the target better following a high (*M* = 2.18) than following a low standard (*M* = 2.54), *F*(1,158) = 4.60, *p* < .05, η^2 = .03. The

expected interaction did not appear ($F < 1.1$). There were no other significant effects (all Fs < 1.9).

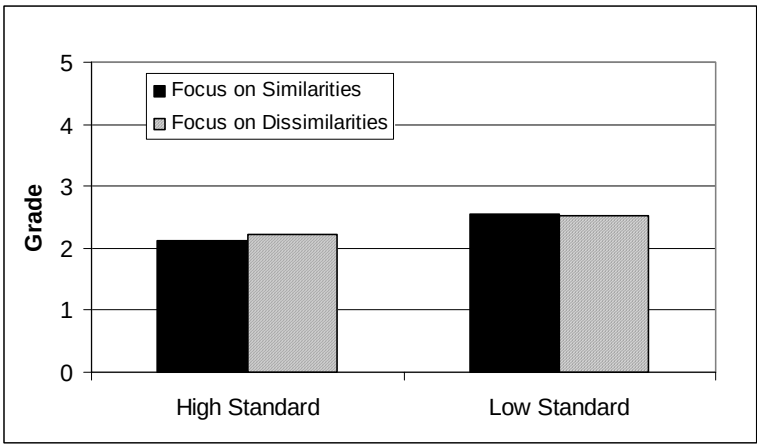


Figure 2.6. Performance judgments of the second pupil (target)

As a second test of the main hypothesis, I re-ran the analysis with the relative ratings as dependent variable. An interaction pattern was expected, but only a trivial main effect emerged, showing that the performance of the second pupil was judged worse[36] than the first pupil in the high standard condition ($M = 3.37$). In the low standard condition, the performance of the second pupil was judged better than the first pupil ($M = 4.98$), $F(1,158) = 63.25$, $p < .001$, $\eta^2 = .29$. This main effect can be seen as an additional manipulation check. There was also a main effect for the factor 'protocols' ($F(1,158) = 3.92$, $p < .05$, $\eta^2 = .02$), meaning that participants were giving worse relative ratings for the target when judging the written protocols ($M = 3.81$) than when judging the audio recordings ($M = 4.34$). There were no other significant effects (all Fs < 1.3).

A third dependent variable was the judgment of the ability of all pupils in Germany. In general, participants judged the ability of the pupils in Germany concerning general knowledge to be rather poor[37] ($M = 3.43$). Especially in comparison with the overall judgment of the first ($M = 2.55$) and the second pupil ($M = 2.36$), the judgment of pupils in general was worse[38], $t(115) = 9.51$, $p < .001$.

To test whether the independent variables also influenced the judgments of all pupils in Germany, the same analysis as before, excluding the factor 'protocols', was

[36] On scale from 1 (*a lot worse*) to 7 (*much better*)
[37] The grading system common in German schools and universities was used. The grades range from 1 to 5 and lower numbers correspond with better grades.
[38] Because participants in study one did not judge the ability of all pupils in Germany only the participants of the second and third study could be used for these analyses.

run with the accordant judgment as dependent variable. It was expected that the biases in the judgments of the two pupils would carry over to the judgments of all pupils in Germany. Participants gave worse grades when focusing on similarities (M = 3.56) than when focusing on dissimilarities (M = 3.29), $F(1,109) = 4.76$, $p < .05$, $\eta^2 = .04$. There was also a tendency for a contrast effects in the judgments: Participants tended to give better grades after low standard (M = 3.35) than after a high standard (M = 3.51), $F(1,109) = 2.03$, $p < .16$, $\eta^2 = .02$. The other effects were not significant (all Fs < 1).

2.5 General discussion of experiments 1-3

When teachers are grading exams, it is a question of fairness that these judgments should be as objective as possible and reflect the actual performance of the pupils. Yet in the case of sequential judgments, it seems likely that the performance of the first pupil would have a systematic and unwanted influence on the performance judgments of the following pupils. It was expected that the performance of the first pupil would be used as a comparison standard for the following targets. The SAM (Mussweiler, 2003a) states that comparisons of this kind are underlying any given judgment task. The SAM assumes that whenever we judge a target, we compare it with a standard, resulting in a relative rather than an absolute judgment of the target. Starting from the idea that the target and the standard can either be seen as similar or dissimilar, the model assumes that this similarity- or dissimilarity-hypothesis will be tested by comparing the standard with the target. This comparison should influence the judgment of the target: If one focuses on similarities, this should lead to assimilation effects and if one focuses on dissimilarities, this should lead to contrast effects. The assumptions of this model were tested in three studies in an complex judgment domain, i.e., in the situation of grading exams. Participants were in the role of a teacher and had to grade exams of two pupils in a sequence. In study one they were given written protocols, in the other two studies they had to base their judgments on audio recordings of exams. The focus of participants was manipulated to be either on similarities or on dissimilarities. This was achieved by using a procedural priming task prior to the judgment task. Additionally, the performance of the first pupil (standard) was manipulated to be either good (high standard) or bad (low standard). Following a high standard, it was expected that participants would judge the performance of the second pupil (target) to be better when focusing on similarities than when focusing on dissimilarities. Following a low standard, it was expected that participants would judge the performance of the second pupil (target) to be worse when focusing on similarities than when focusing on dissimilarities.

What can we conclude from the results of the first three studies? The first and most important finding is that the overall pattern of results seems to justify the main assumption that prior judgments systematically influence subsequent judgments when grading exams in a sequence. The reason for this is that both performances are compared to each other. This is also one of the core assumptions of the SAM. The model predicts that both, assimilation and contrast effects can be caused by the comparison of target and standard. In the studies reported here, it seems that overall

assimilation effects were more likely than contrast effects to appear in the judgments of the second exam (target). Although this pattern of results was not expected, it does not necessarily contradict the SAM. The most parsimonious explanation for the results would be that most participants were following a similarity hypothesis and this led to the assimilation effects in the judgments. Therefore, it seems that the manipulation of the focus via the priming procedure did not have the expected effects. Yet, why was this the case? In the studies reported here, I used the very same priming procedure as in other studies testing the SAM (e.g. Mussweiler, 2001b). By giving participants the unrelated task (prior the judgments) to compare two sketches and to either find all the similarities or differences between them, the respective focus should have become proceduralized (Smith, 1994). One critical aspect of the priming procedure to affect the following judgment and guide the process of comparison might be time. In most studies testing the SAM, the comparison process starts immediately after the manipulation of the focus. This is especially true for the study by Mussweiler (2001b), using the same priming procedure as in the studies reported here. It is, therefore, possible that the strength of the priming procedure will diminish, if participants do not start with the comparison shortly after being primed. The longer the difference between the manipulation of the focus via the priming procedure and the comparison (of target and standard or a different comparison task), the smaller the impact of the priming procedure will be. Two findings support this view: In the first study, the priming manipulation worked fine at the beginning of the study. Even so it was not the focus of this study, assimilation and contrast effects in the expected direction were already found in the judgments of the first pupil (standard). The same effects were found in the judgments of the second pupil (target), but they were smaller. An explanation for the reduced effect might be that the effect of the priming procedure was wearing off over time. The second finding supporting this view is the difference between the results of the first and the second study: In the second study, the materials were presented as audio recordings compared to the written protocols used in the first study. In this second study, I only found assimilation effects in the judgments of the second pupil, i.e., the priming procedure did not affect the judgments as expected. Beside the obvious difference between study one and study two that participants either had to read or to listen to the material, one very important difference was the time to listen or read. Listening to the audio recordings took participants a lot longer than reading through the written protocols. There was also a gap between the first and the second exam when using the audio recordings because the tape had to be fast-forwarded for

some time. Again, an explanation for the priming manipulation not to affect the judgments in the second study might be that the effect had already worn off.

But why did I overall find assimilation effects when the manipulation of the focus did not work properly? At least two explanations might explain this finding. On the one hand, it does make sense that a comparison process makes a focus on similarities more likely. Festinger (1954) as the 'father' of social comparisons already states in his theory that we compare ourselves with others who are similar to us in relevant ways (e.g. Miller, 1984; Wheeler, Koestner, & Driver, 1982). More recently, Kirsten Ruys (2004) proposed in her *Frame-and-compare evaluation* (FACE) model of affective person judgment the idea that objects need to be similar to be comparable to each other. In her approach, she combines the SAM and the *Interpretation/Comparison Model* (Stapel & Koomen, 2001a). It seems that the concepts of comparison and similarity are closely connected. Similarity seems to be a necessary condition for a comparison to occur. However, does a comparison per se also make a focus on similarities more likely? This may, indeed, be the case, as Mussweiler (2001b, p. 501) explains "because comparisons often involve an initial focus on similarities" (see Markman & Gentner, 1996), "similarity testing appears to constitute the default option" (e.g. Chapman & Johnson, 1999; Mussweiler, 2001a, 2003a). On the other hand, apart from being the default option, the comparative focus on similarities was made more likely by the way the two pupils were introduced in my studies. They were of the same age, gender and looked similarly attractive. And above all, both were introduced in the instructions as having been proposed for the same award. So, both pupils must have appeared to be similar to each other, especially on dimensions relevant to the task. Taken together, this made assimilation effects more likely to occur than contrast effects.

Of course, I would also have assimilation effects expected to occur in the third study. Yet, in this case, no biases at all were found in the judgments of the target. The priming procedure was this time located in between the judgments of the two exams. As explained above, the manipulation might, therefore, have been too obvious in this study and made participants de-bias their judgments. A second explanation may also account for these nil results. Maybe in this third study participants did not use the first pupil as a comparison standard. The reason for this might be that the priming procedure not only manipulated the focus of comparison, but also made comparisons per se more likely. Therefore, the first pupil might have appeared as an acceptable comparison standard only in the first two studies. In the third study, participants did probably look for alternative standards to compare the second pupil with. The self (Dunning & Hayes, 1996) as well as best friends

(Mussweiler & Rüter, 2003) are standards that are routinely chosen in a comparison task. Since each participant might have picked an individual standard, this still might have led to assimilation and contrast effects on the individual level, but it also explains the overall nil results when looking at the averaged judgments.

There are additional reasons why the likelihood to find biases in the judgments of exams, as investigated in these studies, should be smaller compared to judgments in other fields. It is obvious to most people that the grading of exams should be as fair (and, therefore, as objective) as possible. Especially for students, fairness in the grading of exams should be a relevant and important issue. For most of the participants in my studies were students themselves, they should have followed this call for objectivity and especially tried to build their judgments only on the individual performance of each pupil. Apart from this, I already discussed two major processes explaining the effects of (positive) hypothesis test-strategies on judgments. Following a positive test-strategy when testing a hypothesis might make the activation of hypothesis consistent information as well as the hypothesis consistent interpretation of information more likely (Trope & Liberman, 1996)[39]. In the studies reported here, only the latter might have occurred, because participants were not actively searching for information, but all received the same pieces of information. Therefore, this might have made a bias in the judgments less likely to occur compared to most social comparison paradigms where participants usually perform a memory-based information search (e.g. Mussweiler & Bodenhausen, 2002). In this case, a large number of information consistent with each hypothesis is assumed to be available in memory (c.f. Hastie & Park, 1986, for a distinction of memory-based and on-line judgments). This assumption is supported by the work of Williams and colleagues (Williams, Cafferty, & DeNisi, 1990) showing that on-line judgments are more accurate than memory-based judgments. Considering these restraints, the results of the present studies seem to be even more notable.

[39] It is important to note that a positive test-strategy has to be distinguished from a confirmatory test-strategy (see Klayman & Ha, 1987)

Chapter 3
-
Comparative judgments:
A closer look at the underlying process of hypothesis testing

3.1 Experiment 4: Grading students taking a psychology exam

3.1.1 Overview

Whereas the last chapter focused on the question whether and in what direction prior judgments have a systematic influence on subsequent judgments, this chapter is concerned with the mediating processes that may explain this influence. The basic idea I started with is that comparisons between prior and subsequent performances may be a central underlying process. This idea is shared by the *Selective-Accessibility Model* (SAM, Mussweiler, 2003a). The concrete assumptions about how comparisons might influence judgments in a sequential judgment situation are derived from this model. The SAM is a recent model that already includes or integrates ideas from earlier models. Apart from the fact that this model sees comparisons as a central process influencing judgments, this model was chosen for two additional reasons. First, the model especially focuses on and makes precise assumptions about the processes underlying comparative judgments. Second, the model is concerned with the selective activation and interpretation of information rather than with the differential use of (context) information. In the case of oral exams teachers are actively searching for information concerning a student's ability or knowledge by asking questions, a situation that was reproduced in the paradigm used in the studies reported in this chapter.

The first three studies reported in the last chapter demonstrate that comparison processes can influence or bias the evaluative performance judgments of 'teachers' grading pupils in a sequential judgment situation. However, these effects were not as strong as expected and only in one (study one) out of three studies the (in terms of the SAM) expected pattern was found, i.e., 'teachers' showed assimilation or contrast effects dependent on the hypothesis they were following (focus on similarities vs. on dissimilarities, respectively). One reason for the weakness or inconsistency of the effects might be that all participants were given the same information to base their judgments on. The SAM suggests assimilation and

contrast effects in judgments to be caused by an underlying process of selective information acquisition (and activation) when comparing target and standard. This process of information acquisition is supposedly guided by a hypothesis. The SAM assumes that participants will employ a certain (positive) test-strategy when testing this hypothesis (cf. Klayman & Ha, 1987). The SAM further assumes that it his hence likely that information consistent with the hypothesis will be activated and also that ambiguous information will be interpreted in line with that hypothesis. In the paradigms used in the first three studies (chapter 2), participants were not actively searching for information, but all received the same kind of information. Therefore, the only source for a bias could have been at the stage of interpreting the information given. An additional important explanation for the weakness of the effects could be the time lag between the manipulation of the focus (on similarities vs. on dissimilarities) and the comparison and judgment, respectively. The effects of the priming procedure, which was used to manipulate the focus of participants, may have weakened or worn off over time. Still, one can conclude from the first three studies that the prior performance judgments influence subsequent ones and these effects are in line with the assumptions of the SAM.

In the following chapter, the main focus is not on the effects of prior judgments on subsequent ones per se, but on the processes underlying these effects. More precisely, the focus in this chapter is on the process of hypothesis testing and the process of information acquisition as processes mediating assimilation and contrast effects in judgments according to the SAM. In the following three studies, participants were again put in the role of a teacher. However, this time they were more actively involved in testing the knowledge of students. In a computer built simulation, participants had to test two (or three) virtual students (or pupils) by asking questions and receiving answers from the students. In the end, they had to grade the performance of these students. Since the process of information acquisition was of special interest, the questions asked by the participants (‘teachers’) were examined. It was expected that participants would be following and testing a certain hypothesis when comparing two students and that the difficulty of the questions asked by participants would vary in accordance with that hypothesis. Since this time participants could actively guide the process of information acquisition, stronger effects compared to the ones found in the first three studies were expected. Since participants would be testing their expectations from the very beginning of the first exam, this should make it more likely for the priming procedure[40] to affect the

[40] The priming procedure was used to manipulate the hypothesis of participants (focus on similarities vs. on dissimilarities).

comparison process and hence the final judgments. Following the first three studies reported in the last chapter, in the fourth study the same two variables were manipulated to test the assumptions of the SAM mentioned above. On the one hand, the quality of the standard was manipulated. The performance of the first virtual student tested by the participants was either good (high standard) or bad (low standard). On the other hand, the hypothesis guiding the comparison process of participants was manipulated: Participants were either primed to focus on similarities or on dissimilarities. In accordance with the SAM, an interaction pattern was expected for the questions chosen for and the judgments given to the second virtual student (target).

The SAM suggests that a comparison of standard and target preceding the final judgment of a target will always be guided by a hypothesis. As explained before (cf. chapter one), there are two general hypotheses people can follow when comparing standard and target. They can either follow a similarity- or a dissimilarity-hypothesis. Depending on the focus of participants that was manipulated to be either on similarities or on dissimilarities and on the performance of the standard (first student), participants should develop different expectancies for the performance of the target (second student). When focusing on similarities, participants should expect a good performance of the target after a high standard and a bad performance after a low standard. When focusing on dissimilarities, participants should expect a bad performance of the target after a high standard and a good performance after a low standard. These expectancies should guide their test strategy, which should be reflected in the kind of questions chosen. More precisely, the performance expectancies held by participants should influence the level of difficulty of the questions chosen. It was predicted that participants expecting a well performing student would chose more difficult questions than participants expecting a badly performing student. Assuming that giving a correct answer to a difficult question is an indicator for a good student and an incorrect answer to an easy question is an indicator for a bad student, this strategy could be called a positive test-strategy in the sense of Klayman and Ha (1987). This can be distinguished from a confirmatory test-strategy (see Snyder & Swann, 1978; Wason, 1960): In the case of a confirmatory test-strategy, participants should ask easy questions when expecting a well performing student and difficult questions when expecting a badly performing student, thus, making the confirmation of their expectancies or hypotheses likely, assuming that the probability to answer an easy question correctly is higher for all students than the probability to answer a difficult question correctly.

At least two explanations may explain why participants should employ a positive test-strategy. On the one hand, the concepts of ‘good academic performance’ and ‘difficult questions’ should be associated for most people. Be it that good students have a higher ability to answer difficult questions or that good students ask more difficult questions, the two concepts should be closely related. Therefore, participants would come up with a hypothesis using a “matching-heurism” (Evans, 1989; Evans & Lynch, 1973). On the other hand, it might be more diagnostic or informative to ask difficult questions when testing the performance of an ostensibly good student.

To understand this idea, one should look back at the assumptions of the SAM: The model states that the process of comparing standard and target is guided by a hypothesis, and that this hypothesis is being tested during the comparison. To be more precise, people test the hypothesis they focus on by searching for information about the target. People do not randomly sample information about the target, but rather try to follow a strategy to test their hypothesis. In many instances – especially if people want to be unbiased as in the case of grading exams – they search for information that seems to be relevant or informative for them (cf. Trope & Liberman, 1996, p. 246). If a teacher expects a student to perform well and if the teacher wants to test this assumption, it seems more informative or diagnostic to ask difficult questions. If the student is not able to answer difficult questions correctly, he cannot be a good student, but if he does, he must be a good student. Easy questions would not be as informative. Of course, a good student should also be able to answer easy questions correctly, but the ability to answer easy questions correctly does not make a good student. Only if the student is not able to answer easy questions correctly, this information is relevant for the hypothesis of the teacher, i.e., the expectancy of the student to perform well. In this case, difficult questions are more diagnostic for testing the hypothesis than easy questions. However, if a teacher expects a student to perform badly and if that teacher wants to test this assumption, easy questions are more informative or diagnostic than difficult questions.

Fiedler and colleagues (Fiedler, Walther, & Nickel, 1999) also express this idea when explaining that the confirmatory test-strategy in the work by Snyder and Swann (1978) is rather a positive test-strategy. In the work by Snyder and Swann, participants were in an interview-like situation and had to test the hypothesis that their interview partner was either extraverted or introverted. Participants selected more questions concerned with extraverted behavior if testing an extraversion-hypothesis; they selected more questions concerned with introverted behavior if testing an introversion-hypothesis. This led participants to an impression of their

interview-partner that was consistent with their hypothesis. The authors, thus, called this one-sided test-strategy a *confirmation bias*. Fiedler and colleagues interpret these findings in light of an adaptive, positive test-strategy:

> Positive testing affords a reasonable strategy in a world that is not strictly dichotomous, so that inverted answers to introversion questions are not as informative about extraversion as answers to direct extraversion questions are. In such a world, in which the negation of an attribute does not imply the assertion of its opposite, positive testing will provide more relevant and diagnostic information than balanced questioning. (p. 6)

For the evaluative performance judgments, the according pattern was expected as in the first three studies. Following a high standard, it was expected that participants would judge the target exams to be better when focusing on similarities (assimilation) than when focusing on dissimilarities (contrast). Following a low standard, participants were expected to judge the target exams to be worse when focusing on similarities (assimilation) than when focusing on dissimilarities (contrast).

In contrast to the first three experiments reported in the last chapter, this time participants were not testing general knowledge of pupils but knowledge relevant to psychology students. The task was to examine the knowledge of two virtual students regarding psychological methods.

3.1.2 Method

Participants

We recruited 66 students at University of Heidelberg as participants for this experiment, for which they received either € 6 or partial course credits in return[41]. Five participants did not follow the instructions while working on the task. The data from these participants was dropped from the analyses. Finally, another five participants were not psychology students. Since it was expected that the lack of knowledge of these students regarding psychological methods would influence their performance judgments, these students were also excluded from the analyses. Summing up, I analyzed the data of 56 (38 female and 9 male) participants[42].

[41] The study was run together with an independent second study. The other study was following this study in the experimental course of events.
[42] Unfortunately, nine participants did not answer the demographic questions.

Participants were between 19 and 29 years old (*M* = 22). All of these participants were students of the field of psychology.

Design

This fourth study followed a 2 (Standard: High vs. low) x 2 (Focus: On similarities vs. on dissimilarities) between participants design. Participants tested the knowledge regarding psychological methods of two students in a virtual exam. The standard (first student) was manipulated by changing the probability of correct and incorrect answers during the exam. The focus of the participants was manipulated using a procedural priming procedure preceding the virtual exam. Dependent variables were the difficulty of questions chosen and the grades given to the two students.

Materials and Pre-tests

Priming Procedure. A procedural priming task was used to get participants into a certain mindset where they should either focus on similarities or dissimilarities in a comparison task. This procedure was adapted from Mussweiler (2001b) and consisted of sketches from two scenes that were taken from Markman and Gentner (1996). In both scenes a person (in the one a man in the other a woman) are setting the table for a Christmas celebration (see Appendix 2.A for the pictures). Participants had to compare these two pictures.

Virtual Exams. Participants were in the role of a teacher testing two students in a virtual exam. This was done using a computer-based simulation. In this simulation, participants could test a student by choosing questions from a list and receiving feedback from the virtual students. The feedback was simply the information whether the student had answered the chosen question correctly or incorrectly. Participants could choose from a list of 20 questions, 10 of which were difficult and 10 of which were easy according to a pre-test (see below). The ability of a virtual student to answer a question correctly was determined by two parameters setting the probabilities to answer an easy or a difficult question correctly, respectively. The probability to answer an easy question correctly was always higher than the one to answer a difficult question correctly. This way, the simulation tried to map real performance of students in an exam situation. After asking 10 questions to one virtual student, participants had to judge the performance of this student on a scale from 4.0 (*very bad*) to 1.0 (*very good*), which is the grading scale used at most German schools and universities and was, therefore, common for participants.

Pre-test. The difficulty of questions was examined in a pre-test. In this pre-test, 23 students participated and judged the difficulty of 32 questions on a scale from 1 (*very easy*) to 7 (*very difficult*). Participants were psychology students and, therefore, from the same population as the participants of the main study. 13 participants were judging questions from the field of cognitive psychology, another 10 participants judged the questions from the field of psychological methods. These two fields were chosen, because each psychology student at the University of Heidelberg had to take the same classes in these two fields in the first year. Therefore, an equal amount of expertise was expected among the participants regarding cognitive psychology and psychological methods. The goal was to find 10 easy and 10 difficult questions for the main study. This was not possible for the questions taken from the field of cognitive psychology. Although a fair amount of the questions was judged to be easy, participants judged none of the questions to be very difficult. The mean difficulty ratings ranged from 1.77 to 4.92 on a scale from 1 (*very easy*) to 7 (*very difficult*). For the questions from the field of psychological methods the mean difficulty ratings ranged from 1.00 to 6.00. The ten most difficult and ten most easy questions were chosen to be used in the main study. The mean ratings of the easy questions ranged from 1.00 to 2.80 (M = 1.86, SD = 0.56) and were judged to be easier as the difficult questions ($t(9) = -15.64$, $p < .001$) that ranged from 4.80 to 6.00 (M = 5.34, SD = 0.83). Additionally, the most difficult of the easy questions was still judged to be easier than the easiest of the difficult questions ($t(9) = -4.05$, $p < .01$). See Appendix 3.A for a list of the questions used in the main study.

Procedure

When participants arrived at the laboratory, they first filled out an informed consent form. They were then given the same two pictures of the priming task used in the previous studies (see 2.1.2). Half of the participants were instructed to find similarities between the two pictures. The other half was instructed to find differences between the two pictures.

After that, participants were informed about a second task, the virtual exams of two students. Participants were told in the instructions that they would be in the role of a teacher and had to test the knowledge of two students regarding psychological methods. The two virtual students were introduced as both being psychology students at the University of Heidelberg. Both students were doing their intermediate examinations. The first student was introduced as Christina studying in her fourth semester doing her third exam of the intermediate examinations. The second student was introduced as Katrin studying in her fourth semester doing her fourth exam of the

intermediate examinations. Each student was introduced with a picture showing a neutral (regarding sympathy) black and white picture of a female face. Participants were then testing each of the two students. They chose 10 questions from a list of 20 questions for each exam. After each questions they received the information whether the student answered this questions correctly or incorrectly. For half of the participants the probability of the first student to answer a question correctly was high (high standard)[43], for the other half the probability was low (low standard)[44]. The probability of the second student (target) to answer a question correctly was intermediate[45] and the same for all participants.

After asking 10 questions and receiving the answers, participants were grading the performance of the respective student. Following the two virtual exams, participants were judging the difficulty of the questions they had chosen before. In the end they also had to indicate their gender, age, and what they were studying. Finally, they were fully debriefed and thanked for their participation by the research assistants running the study. On contrast to the previous three studies, participants did not have to propose one of the students for an exam. This task was excluded to investigate whether comparisons between judgments are still likely without this rather explicit cue in the instruction.

3.1.3 Results

In a preliminary analysis, I found gender of the participants to be correlated with the performance judgments (grades given to the students). Female participants (M = 2.59) gave inferior grades[46] than male participants (M = 2.10) for the second ($t(45) = 1.80$, $p < .08$) but not for the first student (M_{female} = 2.22 vs. M_{male} = 2.72, $t(45)$, -1.22, $p > .22$). However, one has to be cautious with interpreting these effects, because there were far more female than male participants in this study (38 female vs. 9 male participants). Therefore, this variable was not used as a covariate in the following analyses.

As a manipulation check for the manipulation of the standard, I performed a standard x focus analysis of variance (ANOVA) on the performance judgments

[43] The probability to answer an easy questions correctly was set to 90% , the probability to answer a difficult question correctly was set to 80%.
[44] The probability to answer an easy questions correctly was set to 50% , the probability to answer a difficult question correctly was set to 20%.
[45] The probability to answer an easy questions correctly was set to 70% , the probability to answer a difficult question correctly was set to 50% .
[46] The German school grading system was used. The grades range from 1 to 4 and lower numbers correspond with better grades.

(grades) of the first student. As expected, the high standard (*M* = 1.40) was given better grades than the low standard (*M* = 3.49), $F(1,52) = 343.04$, $p < .001$, $\eta^2 = .87$. The judgments also showed an interaction pattern (see figure 3.1), $F(1,54) = 2.94$, $p < .10$, $\eta^2 = .05$. Participants judged the performance of the high standard to be better when focusing on similarities (*M* = 1.27) than when focusing on dissimilarities (*M* = 1.52). For the low standard the opposite pattern emerged. The performance was judged to be better when focusing on dissimilarities (*M* = 3.40) than when focusing on similarities (*M* = 3.54). There were no other significant effects (all *F*s < 1).

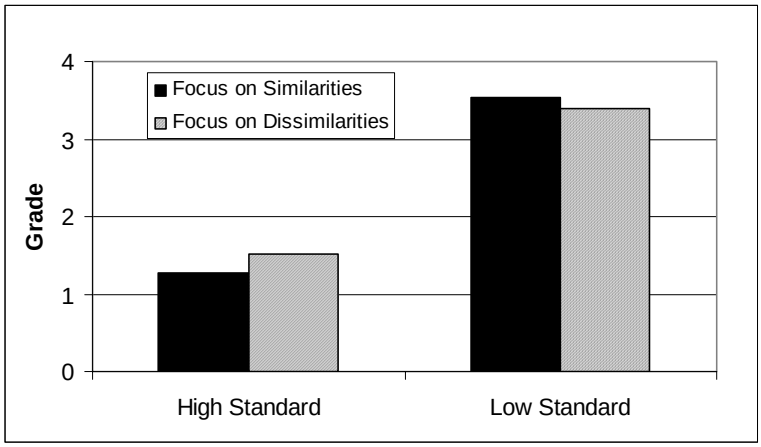


Figure 3.1. Performance judgments of the first pupil (standard)

Difficulty of questions

It was predicted that the difficulty of questions chosen by participants would vary according to their expectations. More precisely, it was predicted that participants would ask more difficult questions when expecting a student to perform well than when expecting a student to perform badly. These expectations should be influenced by the focus of comparison (on similarities vs. on dissimilarities) and the performance of the standard (high standard vs. low standard). There were two measures for the difficulty of the questions. One was based on the 'objective' classification of questions as being easy or difficult based on the pre-test (in the following this will be addressed as 'objective difficulty'). The objective difficulty is the percentage of difficult questions asked in one exam and can, therefore, range from 0 to 100. Each participant also had to judge the difficulty of the questions chosen after the two virtual exams. From these judgments a 'subjective' or individual score for the difficulty of questions was calculated (in the following this will be addressed as 'subjective difficulty'). Participants had to judge the difficulty of the questions on a scale from –25

(*easy*) to +25 (*difficult*). The following analyses will be done using the objective as well as the subjective difficulty of questions as dependent variables.

To test the main hypothesis, I performed a standard x focus analysis of variance (ANOVA) on the difficulty of questions chosen during the exam of the second student (target). There were no significant effects in the two single ANOVAs using either the objective (see table 3.1) or subjective difficulty (see table 3.2) as dependent variable, respectively (all *F*s < 1).

Table 3.1

Difficulty of questions chosen for the second student (target) based on the objective difficulty (percentage of difficult questions)

Standard	Focus on	*M*	*SD*	*N*
High standard	Similarities	44.67	18.85	15
	Dissimilarities	51.25	19.96	16
Low standard	Similarities	44.62	15.61	13
	Dissimilarities	41.67	24.06	12

Table 3.2

Difficulty of questions chosen for the second student (target) based on the subjective difficulty (on a scale from –25 [easy] to +25 [difficult])

Standard	Focus on	*M*	*SD*	*N*
High standard	Similarities	-3.97	7.07	15
	Dissimilarities	-5.83	7.15	16
Low standard	Similarities	-3.44	5.32	13
	Dissimilarities	-3.44	6.85	12

As there were some effects of the independent variables on the grading of the first student, I conducted the same analyses this time with the difficulty of questions asked during the first virtual exam as dependent variables. Participants chose more objectively difficult questions for the high (*M* = 56.77) than for the low standard (*M* = 36.80), $F(1,52) = 18.46$, $p < .001$, $\eta^2 = .26$ (see table 3.3). This can be seen as a proof that participants were at least aware of the different levels of difficulty in the questions. The difference seems to suggest that participants are adjusting the difficulty of questions to the performance of the student much like an adaptive test. There were no other significant effects (all *F*s < 2.1).

Table 3.3

Difficulty of questions chosen for the first student (standard) based on the objective difficulty (percentage of difficult questions)

Standard	Focus on	*M*	*SD*	*N*
High standard	Similarities	54.67	20.66	15
	Dissimilarities	58.75	13.60	16
Low standard	Similarities	32.31	13.63	13
	Dissimilarities	41.67	19.46	12

Using the subjective difficulty as dependent variable, participants chose more difficult questions when examining the high (*M* = -2.67) than when examining the low standard (*M* = -6.8), $F(1,52) = 6.49$, $p < .05$, $\eta^2 = .11$. There was no independent influence for the focus of comparison ($F < 1$). However, the following interaction pattern appeared (see figure 3.2): Participants asked more difficult questions when focusing on similarities (*M* = -1.35) than when focusing on dissimilarities (*M* = -3.91) when testing the high standard. For the low standard the opposite pattern appeared: Participants asked easier questions when focusing on similarities (*M* = -8.60) than when focusing on dissimilarities (*M* = -4.85), $F(1,52) = 3.86$, $p < .06$, $\eta^2 = .07$. This corresponds to the findings of the performance judgments of the standard and will be discussed below.

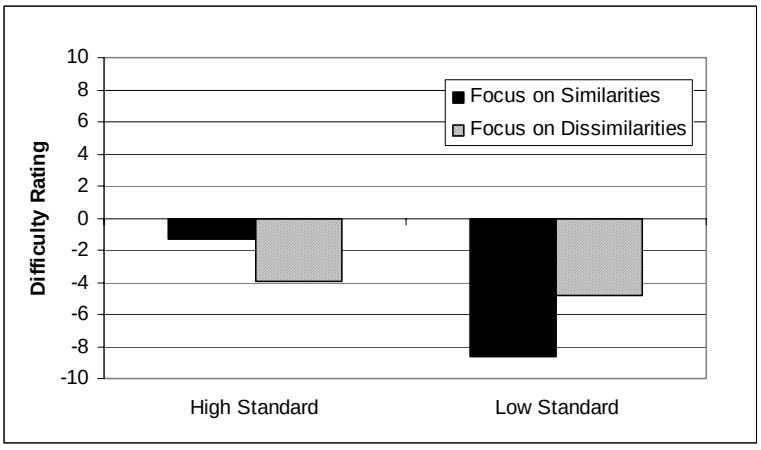


Figure 3.2. Difficulty of questions chosen for the first student (standard) based on the subjective difficulty (on a scale from –25 [*easy*] to +25 [*difficult*])

Performance judgments

It was expected that the performance judgments given by the participants would vary according to their expectations. Participants should give better grades when expecting a student to perform well than when expecting a student to perform badly. These expectations should be influenced by the focus of comparison (on similarities vs. on dissimilarities) and the performance of the standard (high vs. low). To judge the performance, participants used grades of the German grading system common at most schools and universities. The grades ranged from 4.0 (*very bad*) to 1.0 (*very good*). The performance of the virtual students, i.e., the number of correct questions during one exam was dependent on the kind of questions asked. Since each virtual student had two different parameters representing the probability to answer an easy or difficult question correctly and because participants chose different numbers of easy and difficult question, the actual performance of the virtual students differed for each participant. To make the judgments of participants comparable, a relative score was calculated. The number of correct answers given by a virtual student was translated into a score on the grading scale, from 0 % correct answers equaling a 4.0 and 100 % correct answers equaling a 1.0. These transformed scores were then subtracted from the grades given by participants. To make this score easier to interpret, it was finally multiplied with the factor '-1'. In this new scale, representing the 'relative grades', positive scores represent an overrating and negative scores an underrating of the actual performance of the virtual students.

To test the main hypothesis, I performed a standard x focus analysis of variance (ANOVA) on the relative grades given to the second student (target) as dependent variable. There were no significant effects (all *F*s < 1.1), but the pattern was in the expected direction (see figure 3.3). Participants judged the performance of the target after the high standard better when focusing on similarities (*M* = -0.05) than when focusing on dissimilarities (*M* = -0.13). For the low standard the opposite pattern emerged. The performance was judged better when focusing on dissimilarities (*M* = -.12) than when focusing on similarities (*M* = -.13).

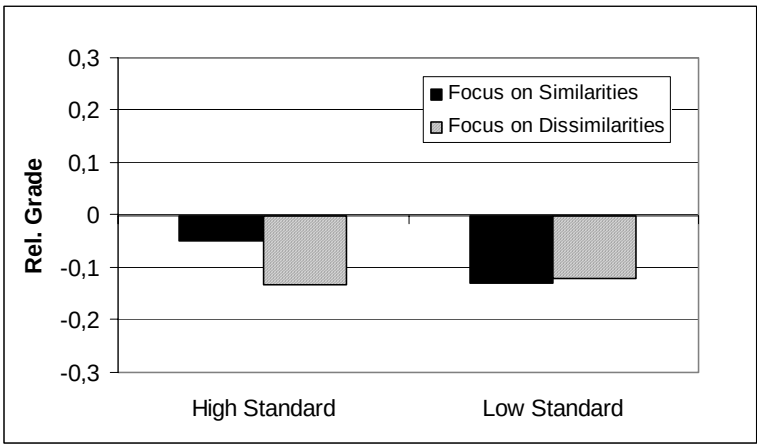


Figure 3.3. Performance judgments of the second student (target) based on relative ratings (positive values represent an overrating, negative values an underrating of the actual performance)

3.1.4 Discussion

The main goal of this fourth study was to take a closer look at the processes underlying comparison effects in sequential evaluative performance judgments. The focus was especially on the process of hypothesis testing that plays a central role in comparison effects according to the SAM (Mussweiler, 2003a). This model explains the processes that lead to assimilation and contrast effects in judgments. Central to the model is a comparison of standard and target prior to the judgment of a target. The model states that people test a certain hypothesis when comparing the target to be judged with a given standard. More precisely, participants follow either a similarity- or a dissimilarity hypothesis. During the comparison process, information will be acquired to test this hypothesis, following a certain (positive) test-strategy. The present study focused on this process of information search or hypothesis testing. Participants were in the role of a teacher and had to grade the performance of two students in an exam regarding psychological methods. To test a student, participants could ask questions and received feedback whether the student answered the questions correctly or incorrectly. When examining the second student (target), participants were expected to compare her with the first student (standard). Two independent variables were manipulated. First, the performance of the first student was manipulated to be either good (high standard) or bad (low standard). Second, the hypothesis of participants was manipulated using a procedural priming task. Participants were primed to focus either on similarities or on dissimilarities. These

two variables should influence the expectancies of participants regarding the performance of the second student (target). The following pattern was predicted: After a high standard participants should expect the target to perform well and, therefore, ask more difficult questions when focusing on similarities than when focusing on dissimilarities. After a low standard the opposite pattern should emerge: Participants should expect the target to perform badly and, therefore, ask more easy questions when they were focusing on similarities than when they were focusing on dissimilarities. Thus, participants were expected to follow a positive (but not necessarily a confirmatory) test-strategy.

Two measures were used to measure the difficulty of the questions asked. The 'objective difficulty' was based on the ratings of a pre-test and the 'subjective difficulty' was based on the individual ratings of the participants. Since the appraisal of a question's difficulty depends on participants' individual level of knowledge and education, the subjective difficulty ratings were expected to be the better measure. However, the results showed no differences between the different conditions in the difficulty of the questions asked to the target, no matter which of the two measures used.

Besides the difficulty of the questions, I was also interested in the performance judgments of the target. Following the first three studies and according to the assumptions of the SAM, assimilation as well as contrast effects were expected in the performance judgments of the target. The SAM states that assimilation effects in judgments are expected when people follow a similarity-hypothesis whereas contrast effects are expected when people follow a dissimilarity-hypothesis. The hypothesis participants followed in this study was manipulated using a procedural priming task that either led participants to focus on similarities or on dissimilarities. However, in accordance with the nil results using the difficulty of the questions as dependent variable, no effects were found in the judgments of the target either. Even so the pattern was in the expected direction, it was not statistically significant.

How can one explain that neither effects for the difficulty of the questions nor for the performance judgments were found? At least two explanations can account for these nil results. First, it might be that one of the manipulations did not work as expected. However, this seems not to be very likely if one considers the questions asked and the grades given to the first student (standard). Participants not only judged the performance of the high standard to be better than the performance of the low standard; they also asked more difficult questions in the high standard condition than in the low standard condition. This proofs that participants were aware of the performance differences of the standard in the two conditions and were also aware of

the different difficulty levels of the questions. Additionally, there was also a tendency for participants to ask more difficult questions and to give better grades to the high standard when focusing on similarities than when focusing on dissimilarities. In the other condition the opposite pattern emerged: Participants asked more easy questions and gave inferior grades to the low standard when focusing on similarities than when focusing on dissimilarities[47]. In other words, assimilations effects appeared when participants were focusing on similarities and contrast effects appeared when participants were focusing on differences. This shows that the priming procedure had the expected effect already at the beginning of the study if one considers the following thoughts: The SAM assumes that every judgment of a target is preceded by a comparison of the target with a standard. Since there was no comparison standard given (in the task) for the first virtual student, participants might have used the performance at the beginning of the exam as a standard. In the high standard condition they would have developed the expectation of that student to be rather good and in the low standard condition to be rather bad. Later in the exam, they might have tested their expectations about that student. This would explain the specific pattern of questions asked and grades given to the first student. However, even if the priming manipulation worked for the first student, its influence might have weakened over time and have been too weak to have an effect during the second exam (cf. chapter two).

A second possible explanation might be that participants did not use the first virtual student as a standard to compare the second student with. The SAM states that there is always a process of comparison preceding the judgment of a target. However, if participants in the present study did not use the first student as a standard, they might have chosen any standard that came to their mind, e.g., their own ability or the ability of an 'average' student. As explained before, they might also have used the performance of the student itself at the beginning of the second exam. For the performance of the second student was average, this would have led to the same expectations for the rest of the exam in all conditions and would explain that there were no differences between the conditions. It is also possible that each participant individually chose a standard, which might have resulted in random effects and this could also explain the nil effects on the group level[48]. This

[47] The described pattern was only true using the subjective difficulty score. As explained before, it was expected that this score would be a more precise measure of the difficulty of questions than the objective difficulty score. This is because there are quite large individual differences in what people judge to be difficult, which might depend on variables such as the individual educational history.

[48] On the individual level, a bias in the judgments in form of an assimilation or contrast effect might still have occurred.

explanation was tested in the following fifth study. To make it sure that participants would compare the two virtual students, they were again told that the two virtual students were competing for an award. Participants were again instructed that they would have to make a proposal in the end which one of the students to give the award to.

3.2 Experiment 5: Comparing students taking a general knowledge exam

3.2.1 Overview

Comparisons between different performances may be the underlying process causing the influence of prior judgments on subsequent ones in a sequential judgment situation. This idea is shared and further developed by the SAM (Mussweiler, 2003a). Following the assumptions of the SAM, both, assimilation and contrast effects may occur when (comparative) judgments have to be made. Whether assimilation or contrast effects occur depends on the hypothesis people follow when making judgments. The model explains these judgmental biases by a comparison process preceding the judgment of a target. During this process the target to be judged is compared with a standard. According to the model, this comparison is done following a (similarity or dissimilarity) hypothesis. This hypothesis is tested during the comparison process and influences what kind of information is searched and activated by the person making a judgment.

In the fourth experiment, this assumption was tested using evaluative performance judgment in exams. This fourth study focused on the difficulty of the questions participants asked when testing the performance of two students. It was expected that the manipulation of the performance of the first student (high vs. low standard) and of the hypothesis participants followed (focus on similarities vs. on dissimilarities) would influence the expectations of the 'teachers' and that this would lead to differences in the questions asked. When participants expected the performance of a student to be good[49], they should test this by asking more difficult questions than when they expected the performance to be bad[50]. Thus, participants were expected to follow a positive – but not necessarily a confirmatory – test-strategy. Although both, the manipulation of the standard (high vs. low) and of the focus (on similarities vs. on dissimilarities), had the expected influence on the questions asked and the grades given to the first student, there were no effects found for the second student. A likely explanation for this might be that participants did not compare the second (target) with the first student (standard). Therefore, in this fifth study the instructions were slightly changed to ensure that participants would

[49] E.g., in the high standard and focus on similarities condition.
[50] E.g., in the low standard and focus on similarities condition.

compare the two students with each other. The task for participants was the same as in study four. They were in the role of a teacher and had to grade two exams. This time, general knowledge rather than psychological methods was the topic of the exams. This was done to make it possible to use students from the general student population as participants. More important, this time participants were again told that the two students were competing for an award and that participants would have to make a proposition who to give the award to in the end.

3.2.2 Method

Participants

We recruited 72 persons at the University of Heidelberg as participants for this experiment, for which they received either € 3 or partial course credits in return. Three participants did not work on the priming task as thoroughly as instructed (less than 1 minute). Another four participants were talking to each other and showing "disturbing behavior"[51]. The data from these participants was dropped from the analyses. Finally, one participant was student of at least one of the academic fields from which the questions in the virtual exams were drawn. Since it was expected that the knowledge of this student would influence his judgments, this student was also excluded from the analyses. Summing up, I analyzed the data of 64 (22 female and 42 male) participants. Participants were between 19 and 50 years old (M = 25). 58 of the participants were students.

Design

This fifth study followed a 2 (Standard: High vs. low) x 2 (Focus: On similarities vs. on dissimilarities) between participants design. Participants tested the knowledge regarding general knowledge of two students in a virtual exam. The standard (first student) was manipulated by changing the probability of correct and incorrect answers during the exam. The focus of the participants was manipulated using a procedural priming task preceding the virtual exam. Dependent variables were the difficulty of questions chosen, the grades given to the two students, and a relative performance measure.

[51] The research assistant (Geoffrey) wrote "Grobes Stören" into the protocols.

Materials & Pre-test

Priming Procedure. The same priming task as in the previous studies was used to get participants into a certain mindset where they should either focus on similarities or dissimilarities in a comparison task. Participants had to compare two pictures and had to find either similarities or differences.

Virtual Exams. The same computer-based paradigm as in the fourth study was used. The only difference was that participants had to test the general knowledge of the virtual students instead of knowledge regarding psychological methods. Therefore, a different set of questions was used.

Pre-test. The difficulty of the questions was pre-tested. In this pre-test, 23 psychology students participated and judged the difficulty of 40 questions on a scale from 1 (*very easy*) to 7 (*very difficult*). Participants were recruited from the same population as the participants of the main study. The questions were from the domains of chemistry, geography, mathematics and physics. The goal was to find 10 easy and 10 difficult questions for the main study. Overall, the difficulty of the questions ranged from 1.00 to 6.20. The ten most difficult and ten most easy questions were chosen. The mean ratings of the easy questions ranged from 1.00 to 1.70 (M = 1.40, SD = 0.40) and were judged as being easier as the mean ratings of the difficult questions ($t(22) = -14.45$, $p < .001$) ranging from 5.13 to 6.20 (M = 5.62, SD = 0.90). Additionally, the most difficult of the easy questions was still judged to be easier than the easiest of the difficult questions ($t(22) = -9.00$, $p < .001$). See Appendix 3.B for a list of the questions used.

Procedure

When participants arrived at the laboratory, they first filled out an informed consent form. They were then given the same two pictures of the priming task as in the previous studies (see 2.1.2). Half of the participants were instructed to find similarities between the two pictures. The other half of the participants was instructed to find differences between the two pictures.

After that, participants were told about a second task, the virtual exams of two students. Participants were told in the instructions that they would be in the role of a teacher and had to test the general knowledge of two students regarding psychological methods. The two virtual students were introduced as both being students competing for an award. Participants were told that they would additionally have to propose one of the students for the award in the end. The first student was introduced as Christina, being 24 years old, and studying in her second year (fourth semester). The second virtual student was introduced as Katrin, being 23 years old,

and also studying in her second year. Each student was introduced with a neutral black and white picture of a female face. Then, participants were testing each of the two students. They chose 10 questions from a list of 20 questions. After each question they received the information whether the student answered this questions correctly or incorrectly. For half of the participants the probability of the first student to answer a question correctly was high (high standard)[52], for the other half the probability was low (low standard)[53]. The probability of the second student (target) to answer a question correctly was intermediate[54] and the same for all participants.

After asking 10 questions participants were grading the performance of the respective student on an absolute scale and additionally on a relative scale. After grading the second student participant had to indicate whom of the two students to propose for the award. Following the two virtual exams, participants were judging the difficulty of all the questions they had chosen before. In the end they also had to indicate their gender, age and what they were studying. Finally they were fully debriefed and thanked for their participation by the research assistants running the study.

3.2.3 Results

In a preliminary analysis, I found the age of the participants to be correlated with one of the dependent variables (difficulty of questions chosen during the first virtual exam). Older Participants asked easier questions[55] than younger participants (r = -.28, p < .05). Therefore, this variable was used as a covariate in the respective analyses. There was no influence of gender on any of the dependent variables.

As a manipulation check, I first looked at the ratings of the participants which one of the two virtual students they proposed for the award. As expected, most participants (77 %) proposed the first student for the award in the high standard condition. In the low standard condition most participants (77 %) proposed the second student for the award.

---

[52] The probability to answer an easy questions correctly was set to 90% , the probability to answer a difficult question correctly was set to 80%.
[53] The probability to answer an easy questions correctly was set to 50% , the probability to answer a difficult question correctly was set to 20%.
[54] The probability to answer an easy questions correctly was set to 70% , the probability to answer a difficult question correctly was set to 50% .
[55] In this study only the analyses based on the individual difficulty ratings (subjective difficulty) are reported because the subjective difficulty ratings proofed to be the more precise measure in the first study.

As a second manipulation check for the manipulation of the standard, I performed a standard x focus analysis of variance (ANOVA) on the performance judgments (grades[56]) of the first student. As expected the high standard (M = 1.64) was given better grades than the low standard (M = 3.10), $F(1,60) = 91.84$, $p < .001$, $\eta^2 = .61$. There were no other significant effects (all Fs < 2.4).

Difficulty of questions

It was expected that the difficulty of questions chosen by participants would vary according to their expectations. It was predicted that participants would ask more difficult questions when expecting a good student than when expecting a bad student. These expectations should be influenced by the focus of comparison (on similarities vs. on dissimilarities) and the performance of the standard (high vs. low).

In contrast to the fourth study, only the results for the individual (subjective) difficulty ratings are reported in this study[57]. These ratings had proven to be the more precise measure in study four. Each participant had to judge the difficulty of the questions chosen after the two virtual exams. From these judgments a 'subjective' or individual score for difficulty of questions was calculated. Participants had to judge the difficulty of questions on a scale from –25 (*easy*) to +25 (*difficult*).

To test the main hypothesis, I performed a standard x focus analysis of variance (ANOVA) on the difficulty of the questions chosen during the exam of the second student (target). In this analysis the expected interaction pattern emerged (see figure 3.4): After a high standard participants asked more difficult questions when focusing on similarities (M = -3.72) than when focusing on dissimilarities (M = -8.34). After a low standard the opposite pattern emerged: Participants asked more difficult questions when focusing on dissimilarities (M = -5.65) than when focusing on similarities (M = -8.93), $F(1,60) = 5.96$, $p < .05$, $\eta^2 = .09$. There were no other significant effects (all Fs < 1). This shows that participants were guided by certain expectancies when testing the second student.

[56] The German school grading system was used. The grades range from 1 to 4 and lower numbers correspond with better grades.
[57] All analyses were also re-ran using the 'objective ratings'. The effects were in the same direction but smaller compared to the 'subjective ratings'.

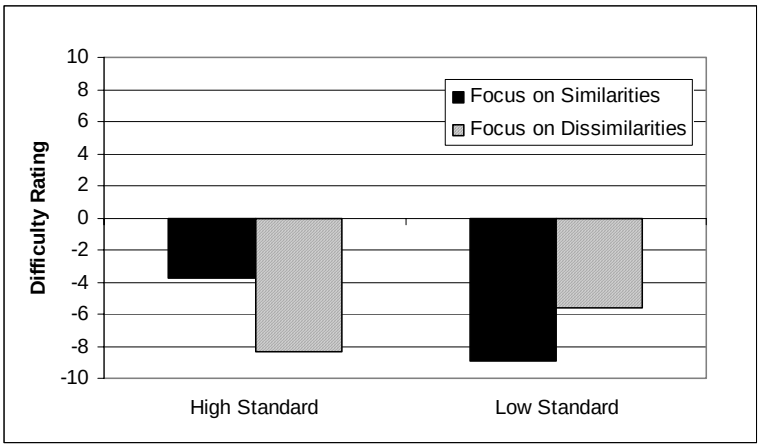


Figure 3.4. Difficulty of questions chosen for the second student (target) based on the subjective difficulty ratings (on a scale from –25 [*easy*] to +25 [*difficult*])

An additional idea was that these expectancy effects should be stronger the less information is available to the person making a judgment. To test this idea, I calculated two difficulty scores, one for the first half of the exam and one for the second half[58]. I re-ran the analyses using both measures, first in two single ANOVAs and second in one ANOVA for repeated measures. The expected interaction was only significant using the first half of the questions as dependent variable, $F(1,60) = 11.08$, $p < .01$, $\eta^2 = .16$. After a high standard participants asked more difficult questions when focusing on similarities ($M = 0.62$) than when focusing on dissimilarities ($M = -9.02$). After a low standard the opposite pattern appeared: Participants asked easier questions when focusing on similarities ($M = -9.64$) than when focusing on dissimilarities ($M = -4.51$).

This interaction was not significant for the second half of the questions ($F < 1.6$), even though the pattern was in the same direction. There were no other significant effects (all Fs < 1). In an additional mixed analysis using the two variables (difficulty during first vs. during second half of the exam) as a within subjects factor there was a significant three-way-interaction, $F(1,60) = 4.89$, $p < .05$, $\eta^2 = .08$. This means that the expectancy effects are stronger at the beginning of the exam when participants have less information compared to the second half of the exam (see figure 3.5).

[58] Since it was expected that strategic thoughts (e.g., “I will start with an easy question to make the student feel more comfortable”) would guide the choices of questions in the very beginning of an exam, the first two questions were dropped.

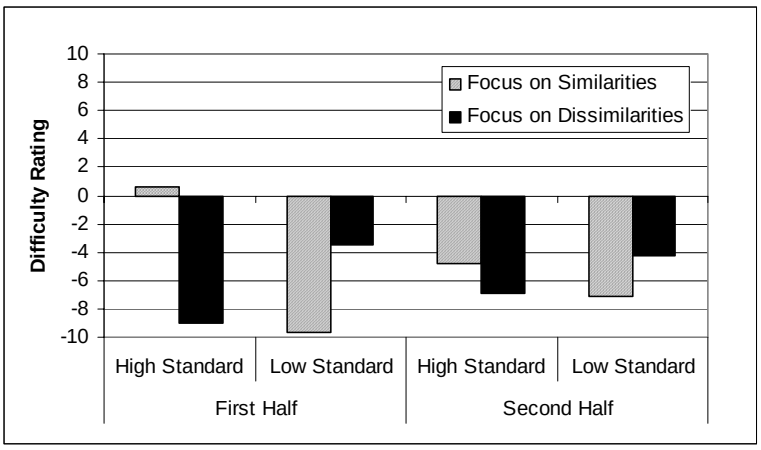


Figure 3.5. Difficulty of questions chosen for the second student (target) during the first vs. the second half of the exam (based on subjective ratings on a scale from –25 [*easy*] to +25 [*difficult*])

Since there were some effects of the independent variables on the grading of the first student, I conducted the same analyses this time with the difficulty of questions asked during the first virtual exam as dependent variables. Participants chose more difficult questions when examining the high (M = -4.38) than when examining the low standard (M = -7.88), $F(1,59) = 5.77$, $p < .05$, $\eta^2 = .09$. This can be seen as a proof that participants were aware of the different levels of difficulty in the questions. The difference seems to suggest that participants are adjusting the difficulty of questions to the performance of the student much like an adaptive test. There was no independent influence of the focus of comparison ($F < 1$). However, the following interaction pattern appeared: Participants asked more difficult questions when focusing on similarities (M = -3.01) than when focusing on dissimilarities (M = -5.45). For the low standard the opposite pattern appeared (see figure 3.6): Participants asked easier questions when focusing on similarities (M = -9.30) than when focusing on dissimilarities (M = -6.46), $F(1,59) = 2.45$, $p < .13$, $\eta^2 = .04$. Examining the simple effects, there was a significant difference between the difficulty of questions asked after a high standard compared to the difficulty of questions asked after a low standard when focusing on similarities ($F(1,59) = 7.45$, $p < .01$, $\eta^2 = .11$). This difference was not significant for the focus on dissimilarities condition ($F < 1$).

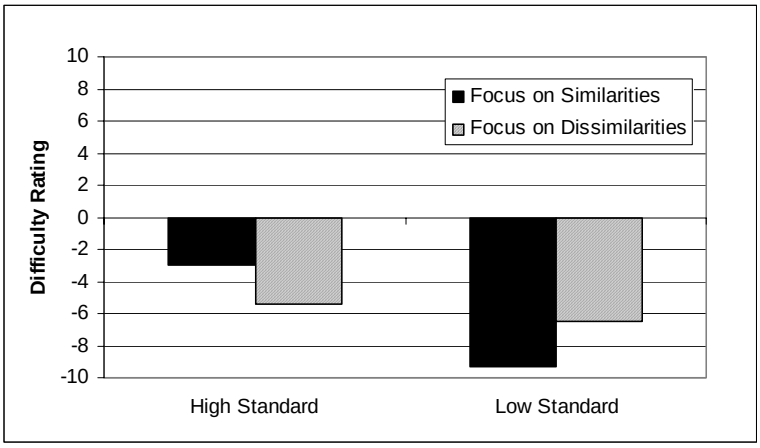


Figure 3.6. Difficulty of questions chosen for the first student (standard) based on the subjective difficulty ratings (on a scale from –25 [*easy*] to +25 [*difficult*])

Performance ratings

As in the fourth study, it was predicted that the performance judgments given by the participants would vary according to their expectations. Participants should give better grades when expecting a student to perform well than when expecting a student to perform badly. These expectations should be influenced by the focus of comparison (on similarities vs. on dissimilarities) and the performance of the standard (high vs. low). To judge the performance, participants used grades of the German grading system common at most schools and universities. The grades ranged from 1.0 (*very good*) to 4.0 (*very bad*). As in the fourth study, the performance of the virtual students, i.e., the number of correct questions during one exam was dependent on the kind of questions asked. Therefore, to make the judgments of participants comparable, a relative score was calculated. The number of correct answers given by a virtual student was translated into a score on the grading scale, from 0 % correct answers equaling a 4.0 and 100 % correct answers equaling a 1.0. These transformed scores were then subtracted from the grades given by participants. To make this score easier to interpret, it was finally multiplied with the factor '-1'. In this new scale, representing the 'relative grades', positive scores represent an overrating and negative scores an underrating of the actual performance of the virtual students.

Additional to the grades, participants also had to judge the relative performance of the target. They had to rate whether the performance of the second student (target) was better or worse than the performance of the first student

(standard). They had to indicate their judgment on a scale from 0 (*a lot worse*) to 30 (*a lot better*).

To test the main hypothesis, I performed a standard x focus analysis of variance (ANOVA) on the relative grades given to the second student (target) as dependent variable. There was an unexpected main effect for the focus of comparison: Participants gave better grades when focusing on similarities (M = -0.03) than when focusing on dissimilarities (M = -0.13), $F(1,60) = 4.59$, $p < .05$, $\eta^2 = .07$. There was also a tendency for an interaction effect suggesting that this difference was bigger after a high than after a low standard as displayed in figure 3.7 ($F < 1.2$). Examining the simple effects, showed the following: After a high standard participants gave significantly better grades when focusing on similarities (M = -0.01) than when focusing on dissimilarities (M = -0.15), $F(1,60) = 5.53$, $p < .05$, $\eta^2 = .08$. This difference (M_{sim} = -0.06 vs. M_{diff} = -0.10) was not significant in the low standard condition ($F < 1$). There was no independent effect for the standard ($F < 1$).

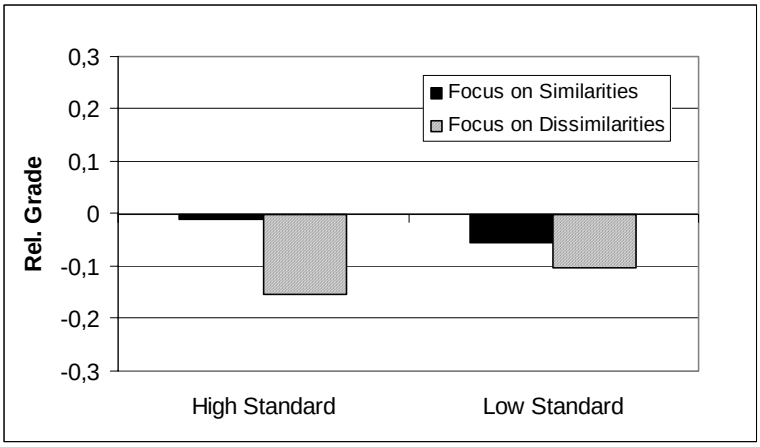


Figure 3.7. Performance judgments of the second student (target) based on relative ratings (positive values represent an overrating, negative values an underrating of the actual performance)

As a second test of the hypothesis, I re-ran the analysis with the relative ratings as dependent variable. An interaction was expected, but only a trivial main effect emerged, showing that the performance of the second student (target) was judged to be worse[59] than the first student (standard) in the high standard condition (M = 9.91). In the low standard condition the performance of the second student was

[59] On scale from 0 (*a lot worse*) to 30 (*much better*).

judged to be better than the performance of the first student (*M* = 20.07), *F*(1,60) = 48.73, *p* < .001, η^2 = .45. This main effect can be seen as an additional manipulation check. There was also an unexpected effect for the focus of comparison: Participants judged the performance of the target to be relatively better when focusing on similarities (*M* = 17.33) than when focusing on dissimilarities (*M* = 12.32), *F*(1,60) = 9.89, *p* < .01, η^2 = .14. There was no significant interaction effect (*F* < 1).

Mediator-Analysis

The SAM assumes that the hypothesis-guided comparison process preceding the judgment of the target is the cause for the bias in the judgment. To test this assumption, three mediator-analyses were calculated using either the main effects of the variables ‘standard’ and ‘focus’ or the interaction effect as the independent variable. The grade was used as dependent variable. It was tested whether the difficulty of questions asked would be a mediator of the effects of the independent variable on the dependent variable.

When using the mean difficulty of questions, the Sobel test statistic was not significant (all test statistics < .90, *p* > .36). However, using the difficulty of questions asked during the first half of the exam, the Sobel test statistic was at least marginally significant for the main effect of the variables ‘standard’ (test statistic = 1.52, *p* < .14) and ‘focus’ (test statistic = 1.50, *p* < .14) as well as for the interaction effect (test statistic = 1.56, *p* < .12).

One has to be cautious with interpreting these findings because the influence of the independent variables on the dependent variable was rather weak. Yet, the results still seem to point in the expected direction, namely that the process of information search (i.e., the difficulty of questions asked) might be the mediating process between the hypothesis (or expectancy) people follow when comparing target and standard and the final judgment of the target.

3.2.4 Discussion

As in the fourth study, the main goal of this fifth study was to take a closer look at the processes underlying assimilation and contrast effects in sequential, evaluative judgments. The focus was on the process of hypothesis testing that plays a central role in comparison effects according to the SAM (Mussweiler, 2003a). The model states that people test a certain hypothesis when comparing the target to be judged with a given standard. More precisely, participants follow either a similarity- or a

dissimilarity-hypothesis. During the comparison process information will be acquired to test this hypothesis. The present study focused on this process of information search or hypothesis testing. Participants were in the role of a teacher and had to grade the performance of two students in an exam testing general knowledge. To test a student, participants had to ask questions and received feedback whether the student answered the questions correctly. While testing the second student (target), it was expected that participants would compare her with the first student (standard). Two independent variables were manipulated. First, the performance of the first student was manipulated to be either good (high standard) or bad (low standard). Second, the hypothesis of participants was manipulated using a procedural priming task. Participants were primed either to focus on similarities or on dissimilarities. These two variables should influence the expectancies of participants regarding the performance of the second student (target). The following pattern was expected: After a high standard participants should expect the target to perform well and, therefore, ask more difficult questions when focusing on similarities than when focusing on dissimilarities; after a low standard the opposite pattern should emerge: Participants should expect the target to perform badly and, therefore, ask more easy questions when focusing on similarities than when focusing on dissimilarities. Thus, it was expected that participants would follow a positive – but not necessarily a confirmatory – test-strategy. In the fourth study this predicted pattern was not found. The reason for this might have been that participants were not using the first student as a standard of comparison. To make sure that participants would compare the second student (target) with the first one (standard), they were told to propose one of the two students for an award after grading the performance of the two.

Participants' subjective ratings were used to measure the difficulty of the questions. This time, the results showed the expected pattern for the subjective difficulty ratings. After a well performing first student (high standard), participants asked more difficult questions when focusing on similarities than when focusing on dissimilarities. However, after a badly performing first student (low standard), participants asked more easy questions when focusing on similarities than when focusing on dissimilarities. This can be explained if one considers participants having and testing different expectations for the target following a positive test-strategy. If they expected a student to be good they asked more difficult questions than when they expected a student to be bad. This positive test-strategy applied by participants seems to be pretty rational or diagnostic. For example, if one wants to test the hypothesis that a student is good, it is more diagnostic to ask difficult question than to ask easy questions. If that student answers the difficult question correctly he or she is

a good student. If that student is not able to answer the difficult question correctly he or she cannot be a good student. Independent of the answer, one will get information either verifying or falsifying the hypothesis. However, to ask easy questions is not as diagnostic in this case. Of course, if that student cannot answers the easy question correctly, he or she cannot be a good student. However, if the student answers the easy question correctly this does not proof that he or she is a good student. Therefore, asking difficult questions seems to be a good strategy if one wants to test the hypothesis that the performance of a student will be good. Accordingly, it is a good strategy to ask easy questions when testing the hypothesis that the performance of a student will be bad (cf. Fiedler et al., 1999, p. 6).

A second interesting finding was concerned with the size of the described expectancy effect over the course of time. It seems reasonable that expectancy effects should be larger when there is less evidence or information available to a person making a judgment. In the paradigm used for this study, participants received 10 pieces of information (answers to questions) for each virtual student. Taken together, all pieces of information would have suggested that the target student was showing an average performance. However, participants were expecting the target student to perform either well or bad, but not average[60]. Therefore, it might happen that they started out to test their expectations but adjusted these to the information they received during the exam. The more information they gathered the less should they be influenced by their initial expectancies. To test this assumption, I looked at the difficulty of questions asked during the first and during the second half of the exam. What I found was the following: The differences in the question difficulty were, indeed, strongest at the beginning of the exam and were getting weaker over time. To be precise, in the end there were no significant differences left between the four conditions. This suggests that even if people do start out with certain expectancies, they are not blind to feedback from the environment. If the information they collect contradicts their expectancies, they are able to adjust to this. This might also explain why the effects in the performance judgments of the targets are rather weak. Only in the high standard condition the expected effects were found. Here, participants judged the performance to be better when focusing on similarities (assimilation) than when focusing on dissimilarities (contrast). For the low standard condition there were no differences. If one assumes that the influence of the expectancies gets weaker the more information is gathered, no strong expectancy effects are expected for the final judgments. Considering the small expectancy effects in the judgments it is

[60] At least, this is what I assume; the expectancies of the participants were not measured in this study

noteworthy that the mediator-analysis still suggests that the strategy to gather information seems to be a variable mediating the effects of the expectancies on the judgments.

There are, of course, alternative explanations why the effects found for the difficulty of the questions were not as strong for the final performance judgments. Two additional reasons should be discussed here. As I have already explained above, the probability for expectancies to affect judgments is higher when there are only few bits of information available to the person making the judgment. Yet, even when information is provided, there is still enough room for an expectancy bias when the information is ambiguous. If the information received is ambiguous, it may be encoded in terms of what is expected (Higgins & King, 1981; Srull & Wyer, 1989). However, in the paradigm used in the present study, participants received as feedback only whether the virtual student answered a question correctly or incorrectly. There was no room for interpretation for this kind of information. This might, therefore, be a reason for the effects in the judgments to be rather weak.

A second explanation is hidden in the architecture of the computer-based paradigm. In the classic studies concerning hypothesis testing (e.g. Snyder & Swann, 1978), it was found that participants will systematically formulate *confirmatory strategies* for testing a hypothesis. In these studies, participants who had to test whether a second person was extraverted asked questions that caused that person to provide actual behavioral confirmation of the hypothesis being tested. However, whether a test-strategy can be called confirmatory depends on the real value of the tested variable. Klayman and Ha (1987) distinguish in this regard a *positive* from a *confirmatory test-strategy* and point out that a positive test-strategy will not automatically lead to confirmatory evidence. One obvious reason for this is that “hypothesis-consistent queries do not preclude negative answers that indicate the absence of the hypothesis-consistent evidence” (Trope & Liberman, 1996, p. 244). The computer paradigm of the present study used two parameters for each virtual student to determine the probability to answer an easy and a difficult question correctly. To mirror reality, the probability to answer a difficult question correctly was always lower than the probability to answer an easy question correctly. Therefore, the performance of each virtual student depended on the number of difficult and easy questions asked by the participants. Now, recall the strategies participants employed: When they expected a student to be good they asked more difficult questions than when they expected a student to be bad. This led to a worse performance of the target in the paradigm. Therefore, the test-strategy employed may be called positive

but it cannot be called confirmatory at all because it made disconfirming evidence more likely to appear.

Therefore, in the following sixth study the paradigm was slightly changed. On the one hand, the information given to participants, i.e., the answers given by the virtual students were real answers that had to be interpreted as being either correct or incorrect. This was done to make it possible for participants to interpret the answers in a hypothesis-consistent way. On the other hand, for each virtual student there was only one general – instead of two – parameters to answer a question correctly. To be more precise, instead of probabilities I used a fix number of correct and incorrect answers for each virtual student. This was done to make the information at least less disconfirming compared to the present study. A second advantage is that the performance of the target was now the same for each participant, independent of the kinds of questions asked[61].

When reanalyzing the data using the objective difficulty of the questions (based on mean ratings from a pre-test) the pattern was the same as for the analyses with subjective difficulty ratings. Yet, even so the pattern was in the same direction, the effects were weaker and not significant for most of the analyses. This is additional proof for the idea that the subjective score is the more exact measure for measuring the difficulty of the questions. There seems to be more error variance using the objective ratings because there are quite significant inter-individual differences when it comes to rate a question to be easy or difficult (cf. Schnotz, 1971). This is a rather interesting finding if one considers how the questions for a show like ‘Who wants to be a millionaire’ or for the final exams of the ‘Abitur’ in Germany are chosen. In these two examples the difficulty of questions is a crucial variable, of course, for very different reasons.

Finally, there was also a minor finding that was not expected. The priming manipulation did appear to have an unexpected influence on the performance judgments of the target. Participants judged the performance of the target to be better when focusing on similarities than when focusing on dissimilarities. A possible explanation for this effect might be that some participants did not use the first virtual student as a comparison standard. If one assumes that some participants chose a high standard independent of the one presented in the study this might explain the results. In this case, a focus on similarities should have led to assimilation and a focus in dissimilarities to contrast effects. Since in a judgment situation there is often more than one available standard, people often rely on routine comparison standards

[61] This is not the case, if participants take the diagnosticity and the difficulty of the questions they ask into account. Yet, it still seems more justifiable to compare the judgments of participants in this case.

(Mussweiler & Rüter, 2003) during the comparison process. One very prominent routine comparison standard used when evaluating other persons is the self (Dunning & Hayes, 1996). Most university students will have a rather positive view of themselves when it comes to general knowledge and this would make the self a high comparison standard. It seems, therefore, possible that (at least some of the) participants in the present study used the self as a comparison standard and this would explain the described effect. To prevent participants from using the self as a comparison standard, in the following study virtual *pupils* instead of *students* had to be examined by participants. This should make the target appear to be less similar to the judge and, therefore, less likely to compare the target with the self, because similarity is a crucial factor when looking for a comparison standard (cf. Festinger, 1954).

To conclude, it seems that the change in the paradigm between the fourth and the fifth study had the expected effect. Making participants compare the two students they were examining led to the expected differences in the test-strategies employed: When expecting a student to be good, participants asked more difficult questions than when expecting a student to be bad, thus, following a positive – but not necessarily a confirmatory – test-strategy. However, these different strategies did not affect the evaluative performance judgments as strongly as expected. The reason for this might be found in the paradigm used that did not leave enough room for interpreting the information collected in an hypothesis-confirming way and even made hypothesis-disconfirming evidence more likely to appear than confirming evidence. Therefore, in the following sixth study the paradigm was adjusted to account for this problem. Finally in the next study, three instead of two virtual students had to be examined. It was expected that the first student as the most prominent one would function as a comparison standard for both following students.

3.3 Experiment 6: Testing the general knowledge of three pupils

3.3.1 Overview

As in the fourth and fifth study, the focus of this sixth study was on the processes underlying assimilation and contrast effects in sequential judgments. More precisely, this study also focused on the process of hypothesis testing underlying comparison effects according to the SAM (Mussweiler, 2003a). This model explains judgmental biases with an underlying comparison process preceding the judgment of a target. According to this model, the comparison of the target with a standard is done following either a similarity- or dissimilarity-hypothesis. This hypothesis is tested – using a positive test-strategy – during the comparison process and influences what kind of information is searched and activated by the person making the judgment.

In the fourth and fifth experiment this assumption was tested for the evaluative performance judgments of exams. The focus was on the questions participants asked when testing the performance of two students. It was predicted that the manipulation of the performance of the first student (high vs. low standard) and of the hypothesis participants followed (focus on similarities vs. on dissimilarities) would influence the expectations of the 'teachers' and this would lead to differences in the questions asked. As predicted, when participants expected the performance of a student to be good[62], they tested this by asking more difficult questions than when they expected the performance to be bad[63]. Thus, participants were following a positive – but not necessarily a confirmatory – test-strategy. In accordance with the SAM, these effects do only occur when the target is compared with a given standard as shown in the fifth study.

Although I found the expected effects in the difficulty of the questions chosen by participants, the effects on the judgments were weaker than expected. An obvious reason for this might be that the information given to participants (i.e., the answers of the virtual students) was contradicting the hypothesis they were testing and could not easily be interpreted in a hypothesis confirming way. Therefore, in this sixth study the paradigm was changed to make the information more open for interpretation and, therefore, make assimilation and contrast effects in the evaluative judgments more likely to appear.

[62] E.g., in the high standard and focus on similarities condition
[63] E.g., in the low standard and focus on similarities condition

The task for participants in the present study was the same as in the fifth study. They were in the role of a teacher and had to grade exams testing general knowledge. In contrast to the fourth and fifth study, the answers of the candidates were realistic answers in this study. This was done to leave more room for interpretation. Additionally, this time participants had to examine pupils instead of students. This was done to prevent participants from comparing the performance of the candidates with their own abilities, i.e., using the self as a comparison standard. Participants were told that they had to rank-order the examined pupils in the end according to their performances. Additionally, to investigate whether the comparison effects would carry on in a sequence longer than two, participants had to examine three instead of two candidates this time.

3.3.2 Method

Participants

We recruited 80 persons at the University of Heidelberg as participants for this experiment, for which they received either € 4 or partial course credits in return. Two participants did not work on the priming task as instructed. The data from these participants was dropped from the analyses. Additionally, another three participants were students of at least one of the academic fields from which the questions in the virtual exams were drawn. Since it was expected that the knowledge of these students would influence their judgments, these students were also excluded from the analyses. Summing up, I analyzed the data of 75 (39 female and 33 male) participants[64]. Participants were between 20 and 36 years old (M = 23). All of the participants were students studying in their 1st to 14th semester (M = 4).

Design

As the previous studies, this sixth study also followed a 2 (Standard: High vs. low) x 2 (Focus: On similarities vs. on dissimilarities) between participants design. Participants tested the knowledge regarding general knowledge of three pupils in a virtual exam. The standard (first pupil) was manipulated by changing the number of correct and incorrect answers during the exam. The focus of the participants was manipulated using a procedural priming task preceding the virtual exam. Dependent variables were the difficulty of questions chosen and the grades given to the three pupils.

[64] Three participants did not give any demographic information about themselves.

Materials

Priming Procedure. Nearly the same priming procedure as in the fourth and fifth study was used to put participants into a certain mindset where they should either focus on similarities or dissimilarities in a comparison task. This time, participants had to compare three pairs of two pictures and had to find either similarities or differences (see Appendix 3.C for the pictures).

Virtual Exams. The same computer-based paradigm as in the fifth study was used. However, there were a couple of minor changes compared to the last study. Instead of setting two ability parameters for each virtual pupil (probability to answer an easy and a difficult question correctly, respectively), this time there was only one parameter for each virtual pupil to answer any kind of question correctly. This parameter was set by using the same random sequence of correct and incorrect answers for each virtual pupil (see Appendix 3.D for the sequences). The parameters were set in a way that the high standard answered eight of the ten (80%) questions correctly, the low standard two of the ten questions (20%) and the two targets (second and third pupil) five of the ten questions (50%). As explained above, this time realistic answers were given to each question instead of just providing the information whether a pupil answered a question correctly or incorrectly. To leave more room for interpretation, the quality of the answer was manipulated for each virtual pupil. Apart from being correct or incorrect, the answers differed in their length, fluency, and the quality of the speech. The latter two attributes were manipulated by the number of stumbles (“mmh”) and the number of inappropriate phrases (e.g., “I did not know we had to prepare this topic.”) included into an answer (see Appendix 3.E for a list of the phrases used). For the high standard the answers were between 70 to 80 words long and there were no stumbles or inappropriate phrases. For the low standard the answers were between 20 to 30 words long and there were four stumbles in each questions and inappropriate phrases in eight of the ten answers. For both target pupils the answers were between 40 to 50 words long and there were two stumbles in each question and inappropriate phrases in eight of the ten answers (see Appendix 3.G for the full material used). Additionally, attached to each answer was either the word “(correct)” or “(incorrect)”. This way not everything was left to interpretation. To ensure that the perception of the difficulty of the questions was similar for the participants, this time a difficulty score[65] for each question was

[65] This score was based on the difficulty ratings of the pre-test used for the fifth study.

attached to each question ranging from 1 (*very easy*) to 10 (*very difficult*) (see Appendix 3.F).

Procedure

When participants arrived at the laboratory, they first filled out an informed consent form. They were then given the three sets of two pictures of the priming task. As in the previous studies (see 2.1.2), half of the participants were instructed to find similarities between the two pictures of each set. The other half of the participants was instructed to find differences between the two pictures of each set.

After that participants were told about a second task, the virtual exams of three pupils. Participants were told in the instructions that they would be in the role of a teacher and had to test the general knowledge of three pupils. Participants were told that they would additionally have to rank order participants based on their performance in the end. This was done to ensure that participants would compare the performance of the pupils with each other. The first pupil was introduced as Christina, the second as Katrin and the third as Susanne. Each pupil was introduced with a neutral black and white picture of the face of the respective pupil. Participants were then testing each of the three pupils. They chose 10 questions from a list of 20 questions. After each question they received the answer plus the information whether the student answered this question correctly or incorrectly. For half of the participants the probability of the first pupil to answer a question correctly was high (high standard: 80%), for the other half the probability was low (low standard: 20%). The probability of the second and third pupil (target 1 and 2) to answer a question correctly was intermediate (50 %) and the same for all participants.

After asking 10 questions, participants were grading the performance of the respective pupil on a scale ranging from 0 (*very bad*) to 50 (*very good*). They also had to judge the performance on five additional scales measuring detailedness, fluidity, comprehensibility, precision, and adequacy of the answers (also on a scale form 0 to 50). Following the three virtual exams, participants were judging the difficulty of the questions they had chosen before (subjective difficulty ratings). After that, participants had to rank order the performance of the three pupils. Then they had to judge the ability of all pupils in Germany as well as their own ability regarding general knowledge. In the end they also had to indicate their gender, age and what they were studying. Finally, they were fully debriefed and thanked for their participation by the research assistants running the study.

3.3.3 Results

In a preliminary analysis, I found no correlation between the age and the gender of the participants with the dependent variables. Therefore, no demographic variables were used as covariates in the following analyses.

As a manipulation check, I first looked at the rank orders of the virtual pupils as suggested by participants. As expected, most participants (92 %) ranked the first pupil highest in the high standard condition. In the low standard condition all participants ranked the second or the third pupil highest.

As a second manipulation check for the manipulation of the standard, I performed a standard x focus multivariate analysis of variance (MANOVA) on the performance judgments[66] of the first pupil. As expected, the performance of the high standard (M = 37.85) was judged to be better than the performance of the low standard (M = 14.85), $F(6,66) = 48.94$, $p < .001$, $\eta^2 = .82$. There were no other significant effects (all Fs < 1).

Difficulty of questions

As in the fourth and the fifth study, it was predicted that the difficulty of questions chosen by participants would vary according to their expectations. It was predicted that participants would ask more difficult questions when expecting a well performing student than when expecting a badly performing student. These expectations should be influenced by the focus of comparison (on similarities vs. on dissimilarities) and the performance of the standard (high vs. low).

As in the fifth study, only the analyses using the individual (subjective) difficulty ratings for the questions are reported here[67]. These have proven to be the more precise measure in the fourth and fifth study. Each participant had to judge the difficulty of the questions chosen after the two virtual exams. From these judgments a 'subjective' or individual score for difficulty of questions was calculated. Participants had to judge the difficulty of questions on a scale from 0 (*easy*) to 50 (*difficult*).

Second pupil (target 1). *T*o test the main hypothesis, I performed a standard x focus analysis of variance (ANOVA) on the difficulty of questions chosen during the

[66] Judgments were made on six dimensions including an overall performance judgment. The other dimensions were detailedness, fluidity, comprehensibility, precision and adequacy of the answers. All answers were given on scales from 0 (*not …*) to 50 (*very …*).

[67] All analyses were also re-ran using the 'objective ratings'. The effects were in the same direction but smaller compared to the 'subjective ratings'.

exam of the second pupil (target 1). Unexpectedly, there were no significant effects at all (all *F*s < 1), as can be seen in table 3.4.

Table 3.4

Difficulty of questions chosen for the second pupil (target 1) based on the subjective difficulty ratings (ranging from 0 [easy] to 50 [difficult])

Standard	Focus on	*M*	*SD*	*N*
High standard	Similarities	18.94	5.65	20
	Dissimilarities	19.96	7.04	19
Low standard	Similarities	19.66	5.06	17
	Dissimilarities	21.39	4.35	19

The results of the fifth study showed that the expectancy effects were stronger at the beginning of each exam when there was less information available to the judges. Therefore, I calculated two difficulty scores, one for the first half of the exam and one for the second half[68]. I re-ran the analyses using both measures, first in two single ANOVAs and second in an ANOVA for repeated measures. There was a significant interaction in the expected direction only for the first half of the questions, $F(1,71) = 4.64$, $p < .05$, $\eta^2 = .06$. After a high standard participants asked more difficult questions when focusing on similarities ($M = 21.87$) than when focusing on dissimilarities ($M = 17.60$). After a low standard the opposite pattern appeared: Participants asked easier questions when focusing on similarities ($M = 14.76$) than when focusing on dissimilarities ($M = 20.56$). This interaction did not show for the second half of the questions ($F < 1$). In an additional mixed analysis using the two variables (difficulty during first vs. during second half of the exam) as a within subjects factor the three-way-interaction was marginally significant, $F(1,71) = 3.16$, $p < .09$, $\eta^2 = .04$. This means that there is a (marginally significant) difference in the interaction patterns found for the questions in the first and the second half of the exam and suggest that the expectancy effects are stronger at the beginning of the exam when less information is available (see figure 3.8). There were no other significant effects (all *F*s < 1).

[68] Since it was expected that strategic thoughts (e.g., “I will start with an easy question to make the student feel more comfortable”) would guide the choices of questions in the very beginning of an exam, the first two questions were dropped.

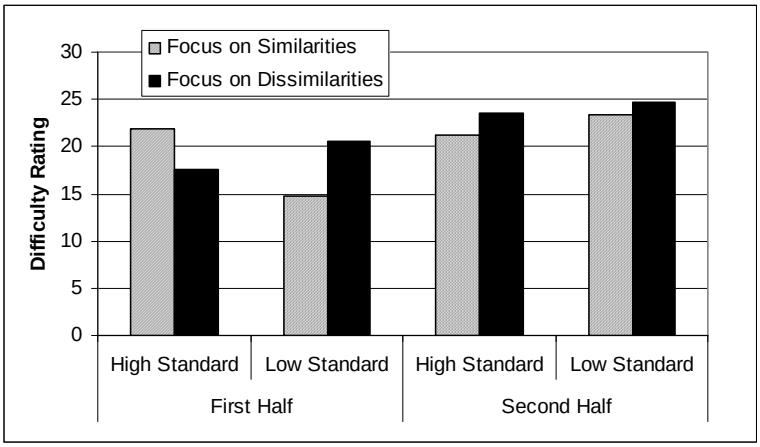


Figure 3.8. Difficulty of questions chosen for the second pupil (target 1) during the first vs. the second half of the exam (based on subjective ratings on a scale from 0 [*easy*] to 50 [*difficult*])

Third pupil (target 2). For the second target, the same effects as for the first target were expected. This would indicate that the first pupil in a sequence is taken as a comparison standard for at least two following pupils. To test the main hypothesis, I performed a standard x focus analysis of variance (ANOVA) on the difficulty of questions chosen during the exam of the third pupil (target 2).

Participants asked more difficult questions after a high (M = 20.39) than after a low standard (M = 16.09), $F(1,71) = 7.79$, $p < .01$, $\eta^2 = .10$. There was also the (marginally significant) interaction in the expected direction (see figure 3.9), $F(1,71) = 3.49$, $p < .07$, $\eta^2 = .05$. After a high standard participants asked more difficult questions when focusing on similarities (M = 21.44) than when focusing on dissimilarities (M = 19.28). After a low standard the opposite pattern appeared: Participants asked easier questions when focusing on similarities (M = 14.14) than when focusing on dissimilarities (M = 17.84). There was no independent effect for the focus of comparison ($F < 1$).

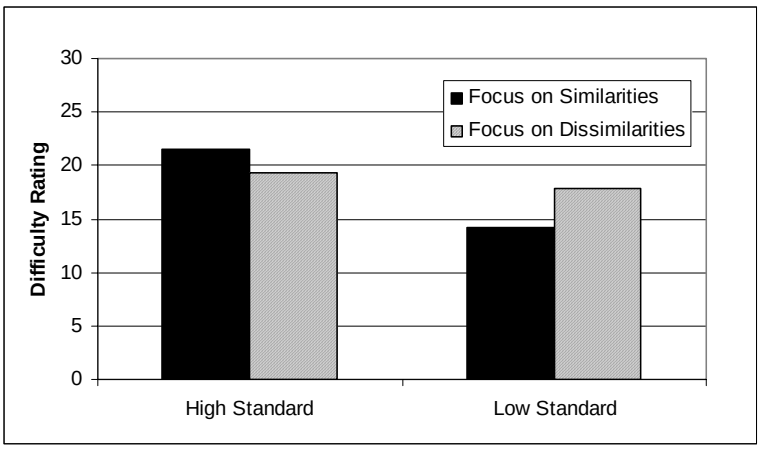


Figure 3.9. Difficulty of questions chosen for the third pupil (target 2) based on subjective ratings (on a scale from 0 [*easy]* to 50 [*difficult*])

To test whether the expectancy effects were again stronger at the beginning of the exam, I calculated two difficulty scores, one for the first half of the exam and one for the second half[69]. I re-ran the analyses using both measures first in two single ANOVAs and second in an ANOVA for repeated measures. Participants asked more difficult questions after a high (M = 20.15) than after a low standard (M = 13.96) during the first half of the exam, $F(1,71) = 9.68$, $p < .01$, $\eta^2 = .12$. This effect was still there for the second half, but weaker and no longer significant ($F < 2$). There was a tendency for the expected interaction as well in the first as in the second half of the exam, but both were non-significant (Fs < 2.3). In an additional mixed analysis using the two variables (difficulty during first vs. during second half of the exam) as a within subjects factor, there was no three-way-interaction ($F < 1$). This means that the expectancy effects were equally strong at the beginning and at the end of the third exam.

First student (standard). Although there were no interaction effects of the independent variables on the grading of the first student, I conducted the same analyses this time with the difficulty of questions asked during the first virtual exam as dependent variables. Participants chose more difficult questions when examining the high (M = 21.26) than when examining the low standard (M = 16.08), $F(1,71) = 14.29$, $p < .001$, $\eta^2 = .17$. There were no other significant effects (all Fs < 1).

[69] Since it was expected that strategic thoughts (e.g., "I will start with an easy question to make the student feel more comfortable") would guide the choices of questions in the very beginning of an exam, the first two questions were dropped.

Performance Judgments

As in the fourth and the fifth study, it was predicted that the performance judgments given by participants would vary according to their expectations. Participants should give better grades when expecting a pupil to perform well than when expecting a pupil to perform badly. These expectations should be influenced by the focus of comparison (on similarities vs. on dissimilarities) and the performance of the standard (high vs. low). To judge the performance, participants this time used a general scale ranging from 0 (*very bad*) to 50 (*very good*). Additionally, participants judged the performance on five more specific scales (measuring detailedness, fluidity, comprehensibility, precision, and adequacy of the answers), also ranging from 0 (*not …*) to 50 (*very …*). Because the performance of the virtual pupils was identical this time for each participant, there was no need to transform the ratings as in the previous studies four and five.

Second pupil (target 1). To test the main hypothesis, I performed a standard x focus multivariate analysis of variance (MANOVA) on the six performance judgments given to the second pupil (target 1) as dependent variable. There was only a contrast effect in the judgments: Participants judged the performance of the second pupil to be better after a low (M = 31.18) than after a high standard (M = 19.80), $F(6,66) = 11.21$, $p < .001$, $\eta^2 = .51$. There were no other significant effects (all Fs < 1.5). Reanalyzing the data in single ANOVAs, using each of the single performance judgments as dependent variables, showed the same pattern in all but one case. Interestingly, there was a significant interaction in the expected direction for the variable ‘detailedness’ (see figure 3.10), $F(1,71) = 6.82$, $p < .05$, $\eta^2 = .09$.

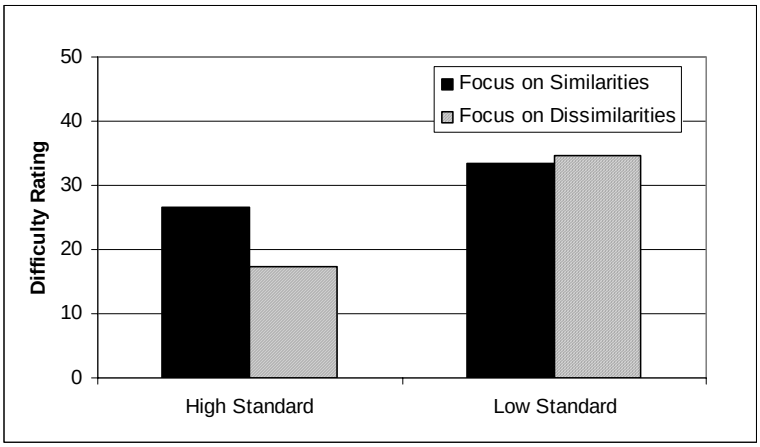


Figure 3.10. Performance judgments of the second pupil (target 1) on a scale ranging from 0 (*very bad*) to 50 (*very good*)

Third pupil (target 2). To test the main hypothesis, I performed a standard x focus multivariate analysis of variance (MANOVA) on the six performance judgments given to the third pupil (target 2) as dependent variable. Again, there was only a contrast effect in the judgments: Participants judged the performance of the third pupil to be better after a low (M = 28.66) than after a high standard (M = 18.45), $F(6,66) = 7.07$, $p < .001$, $\eta^2 = .39$. There were no other significant effects (all Fs < 1.1). Reanalyzing the data in single ANOVAs, using the single performance judgments as dependent variables, showed the same pattern in all but one case. As for the first target, there was a significant interaction in the expected direction for the variable 'detailedness', $F(1,71) = 5.46$, $p < .05$, $\eta^2 = .07$.

Ability judgments

As additional variables, participants had to judge the ability of all pupils in Germany and their own ability regarding general knowledge on a scale from 0 (*very bad*) to 9 (*very good*). To see whether the effects found for the judgments of the virtual pupils would also carry over to these variables, two standard x focus analyses of variance (ANOVAs) on the ability judgments were run.

For the ability judgment of all pupils in Germany there were no effects (all Fs < 1). However, in the judgments of participant's own ability, there was the expected interaction pattern (see figure 3.11), $F(1,71) = 5.14$, $p < .05$, $\eta^2 = .07$. After a high standard, participants judged their own ability to be better when focusing on similarities (M = 5.95) than when focusing on dissimilarities (M = 4.67). After a low standard, the opposite pattern emerged: Participants judged their own ability to be worse when focusing on similarities (M = 4.75) than when focusing on dissimilarities (M = 5.32). These results make sense if participants also used the first pupil as a standard to compare their own ability with. The results also show that the primed focus was still guiding the comparison process even after grading the three pupils in their exams. Since this time the persons judging were also the targets to be judged, this is also a nice replication of the results usually found when testing the SAM in the domain of social comparisons.

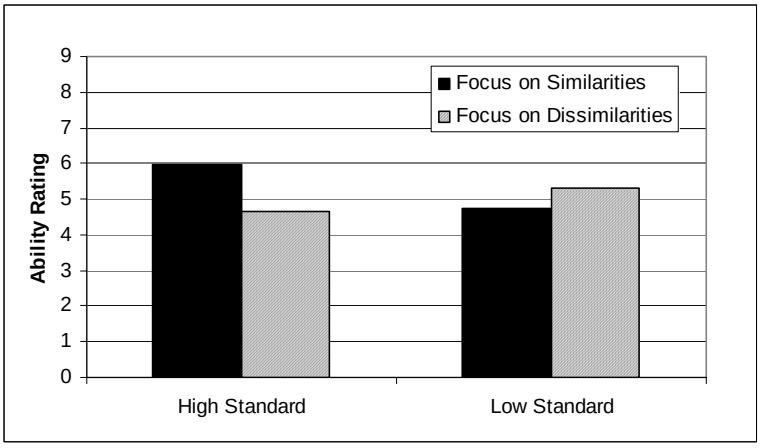


Figure 3.11. Ability judgments of the self, concerning general knowledge on a scale from 0 (*very bad*) to 9 (*very good*)

Mediator-Analysis

The SAM assumes that the hypothesis-guided comparison process preceding the judgment of the target is causing the bias in this judgment. To test this, three mediator-analyses were calculated as in the fifth study (see 3.2.3) for the second as well as for the third virtual pupil using either the main effects of the variables 'standard' and 'focus' or the interaction effect as the independent variable. The mean of the performance judgments was used as dependent variable. It was tested whether the difficulty of questions asked would be a mediator of the effects of the independent variable on the dependent variable.

I first calculated the analyses for the second virtual pupil (first target). Using the mean difficulty of questions, the Sobel test statistic was not significant (all test statistics < .56, $p > .64$). However, when I used the difficulty of questions asked during the first half of the exam, the Sobel test statistic was at least marginally significant for the main effect of the variable 'standard' (test statistic = 1.78, $p < .09$) as well as for the interaction effect (test statistic = 1.74, $p < .09$). I also calculated the same analysis for the third virtual pupil (second target) but I did not find any significant effects (all test statistics < 1.10, $p > .27$).

As explained above (see 3.2.3), one has to be cautious with interpreting these findings because the influence of the independent variables on the dependent variable was rather weak. Still, the findings seem again to point in the expected direction, namely that the process of information search (i.e., the difficulty of questions asked) might be a process mediating between the effects of the hypothesis or expectancy people follow when comparing target and standard, and the judgment of the target.

3.3.4 Discussion

Apart from trying to replicate the findings of the fifth study, this sixth study had two main goals: First, by making the material (i.e., the answers given by the virtual pupils) more ambiguous it was expected to find a stronger bias in the performance judgments of the virtual pupils. Second, it was tested whether the effects found for a sequence of two exams (one standard, one target) could be extended to a sequence of three exams (one standard, two targets).

To see whether the experimentally induced expectancies of participants would translate into different test-strategies, I first focused on the difficulty of questions chosen in the exams[70]. As before, it was predicted that participants would ask more difficult questions when expecting a good performance than when expecting a bad performance. Thus, it was predicted that participants would employ a positive – but not necessarily a confirmatory – test-strategy. On the surface, there seemed to be a no effects for the first target (second pupil). However, the results of the fifth study suggested that expectancy effects should be stronger at the beginning of an exam when there is less information available to the person making the judgment. This was also true for this sixth study. In the first half of the second exam, participants showed the expected interaction pattern: After a high standard they asked more difficult questions when focusing on similarities than when focusing on dissimilarities. After a low standard they asked more difficult questions when focusing on dissimilarities than when focusing on similarities. This means that they used a positive test-strategy to test their expectations of the first target pupil. In the second half of the exam participants showed no such differences in the difficulty of the questions chosen. This underlines very nicely the assumption that expectancies have a stronger influence on information search strategies when there is less information available. This suggests that people do not blindly follow their first expectations but are able to adjust their views to information they receive from the environment.

For the exam of the third pupil, i.e., the second target, participants showed the same test-strategy. They asked more difficult questions when expecting the pupil to perform well than when expecting the pupil to perform badly resulting in the expected interaction pattern. Additionally, there was also a main effect meaning that participants asked more difficult questions after a high than after a low standard. When comparing the questions chosen in the first half with the ones chosen in the

[70] The analyses are based on the (subjective) difficulty ratings given by participants. When reanalyzing the data using the “objective” difficulty scores based on the ratings of the pre-test, the effect were going in the same direction but were smaller.

second half of the exam the following pattern emerged: The main effect was only apparent at the beginning of the exam. The interaction pattern was weaker but still there over the whole course of the exam. So, the effects on the test-strategy seem to be weaker for the second target compared to the first target.

In addition, the influence of new information had a weaker effect on the strategy compared to the exam of the first target. At least two explanations can account for this difference between the two targets. On the one hand, even when participants used the first virtual pupil (standard) as a comparison standard for the second pupil (first target), this does not mean that they had to use the same standard for the third pupil (second target). It seems reasonable that at least some of the participants used a different comparison standard for the second target. They might have used the first target as a comparison standard (as the one directly preceding the second target) or they may have used the average performance of the first two pupils as a comparison standard. Both could explain that the effects were slightly different for the second target. On the other hand, participants might have become bored or distracted over the course of the study. In that case they would not pay too much attention to the information provided in the third exam any more. This could explain why there were only small differences between the questions chosen in the first and second half of the exam. Expectancy effects should become smaller the more information is available. However, this is not the case if the information provided is not perceived and used.

To sum up, the findings of the sixth study regarding the test-strategy employed by participants nicely replicate and extend the findings of the fifth study. Participants ask more difficult questions when expecting a good than when expecting a bad performance, thus, employing a positive test-strategy. This effect is stronger the less information is available. Additionally, it seems that most participants use the first pupil in a sequence as standard for both following pupils.

In the performance judgments I found contrast effects for all variables used but not the expected interaction pattern. Participants judged the performance of both targets to be better after a low than after a high standard. So, it seems that changing the level of ambiguity in the answers of the virtual pupils did not have the expected effect. After all, for each answer the additional information whether the answer was correct or incorrect was provided. It might still be possible to make the material even more ambiguous by leaving out this part of information in future studies. An alternative explanation might be that a focus on dissimilarities resulting in contrast effects might be the default view when comparing a target with a standard. Therefore, it might be difficult to change this view into an assimilative view. Yet, this

could not explain why the manipulation of the focus had the expected effects on the questions chosen and is also not supported by the literature (Chapman & Johnson, 1999; Mussweiler, 2001a, 2003a), but see Stapel & Koomen (2001a).

A different view could be that assimilation and contrast effects are rather independent and influence each kind of comparative judgment to a certain extent. Mussweiler (2003a) suggests that subjective judgments may be the joint product of two independent processes, namely *selective accessibility* and *reference point use*. Whereas the selective accessibility mechanism should make assimilation effects more likely, the opposite is true for using reference points. Therefore, the same comparison may have different underlying tendencies and may even produce assimilation and contrast effects at the same time (e.g. Mussweiler & Strack, 2000b). So, it may be that in the present study participants used their first performance judgment as a reference point and that the resulting contrast effects were relatively stronger than the ones resulting from using selectively accessible information. However, even if these findings are not contradicting the assumptions of the SAM, they can also be explained by the *Interpretation/Comparison Model* (Stapel & Koomen, 2001a) that expects contrast effects only to occur when people are in a comparative mindset. Assimilation effects should only occur when people are in an interpretative mindset.

It should be noted that the expected interaction effects were found consistently for one of the six dependent variables used, namely the judgment of the 'detailedness' of the answers. Since the effect was found on the same variable for both targets, this makes a random effect rather unlikely. It may be that this scale was the fuzziest one, i.e., leaving most room for interpretation, and, therefore, it was most susceptive to be influenced by the expectations of the participants. This variable is at least partial proof for the SAM; even so this model cannot explain why the expected judgment biases were only found on this variable. Additional support for the SAM comes from the mediator-analyses. Considering the mixed findings in the expectancy effects, it is noteworthy that the mediator-analyses still suggests that the strategy to gather information is, indeed, a variable mediating the effects of the expectancies on the performance judgments.

An additional finding worth mentioning is that there was no independent (i.e., main) effect of the focus manipulation on the performance judgments. This was found in the fifth study and was explained by the assumption that some participants might have used themselves as a high comparison standard. To prevent participants from doing this, virtual *pupils* instead of virtual *students* were used this time. Since participants were students themselves, it would have been inadequate to compare

the virtual pupils with themselves. As expected, this time there were no independent effects of the focus manipulation.

In the end, I would also like to mention the interesting finding that the manipulation of the expectancies also affected the self-judgments of participants in a way consistent with the SAM. After testing the general knowledge of three virtual pupils, participants had to judge their own ability regarding general knowledge. In these ratings I found the expected interaction pattern. After a high standard, participants judged their own ability to be better when focusing on similarities than when focusing on dissimilarities. However, after a low standard, participants judged their ability to be better when focusing on dissimilarities than when focusing on similarities. It is astonishing that participants still used the performance of the first pupil as a comparison standard for judging themselves. These results also show that the priming procedure was strong enough to influence the focus of participants over the course of the whole study. Most interestingly, it shows that the expectancy effects suggested by the SAM are found more clearly in self-judgments than in the performance judgments of the virtual pupils. Why is this the case?

As explained before, the activation of information might be a crucial variable. In the case of self-judgments information is searched in and activated from memory. Different kinds of information are activated when following different expectations. Since the process of information search is guided and biased by the expectations held, expectancy-consistent information is more likely to be activated than expectancy-inconsistent information. In the case of testing the virtual pupils in the paradigm used, the same kind of information was given to every participant; independent of the expectancies they followed. This means that all information (consistent and inconsistent) was available to all participants. Expectancy effects were still expected to guide the interpretation of the information to be rather expectancy-consistent. Yet, this should produce weaker effects than only having expectancy-consistent information activated from memory. The difference between the two judgments mirrors the distinction of memory-based and online judgments (Hastie & Park, 1986). This assumption is supported by the work of Williams and colleagues (Williams et al., 1990) showing that on-line judgments are more accurate than memory-based judgments. Although participants could also have tried to discount expectancy-inconsistent information, this seems rather unlikely (see for example Fiedler, 2000).

Summing up, the findings of this sixth study nicely replicate and extent the findings of the fourth and fifth study. Overall, the findings are in line with the assumptions of the SAM considering the processes underlying assimilation and

contrast effects: When people compare a target with a standard they first build and then test expectations using a positive test-strategy. In the studies four to six the difficulty of the questions participants asked as ‘teachers’ during an exam was influenced by their expectancies. An important finding is that the effect of these expectancies is stronger the less information is available. Participants are able to adjust their expectancies to the evidence they find, they are not blinded by the light of their expectancies (maybe only a little short-sighted). This might also explain why the effects in the final evaluative judgments are not as consistent with the SAM as expected. Again, people do not base their judgments solely on their expectations, but reality seems to play an important role as well. Another important aspect is that biases in judgments are much more likely when the process of information acquisition is memory-based. The expected assimilation and contrast effects in the judgments were most clearly found in participants’ self-judgments; in this case, more expectancy-consistent than expectancy-inconsistent information may be searched and activated by participants. In the case of the paradigm used in this sixth study, the performance of the targets to be judged was the same for all participants, i.e., expectancy-consistent as well as expectancy-inconsistent information was available. Participants could still have reinterpreted the inconsistent information to make it more consistent, but reinterpretation should lead to smaller effects than when inconsistent information were not activated at all.

3.4 General discussion of experiments 4-6

Assimilation and contrast effects in judgments are not a new topic in the field of social and cognitive psychology. Assimilation and contrast effects have been found to affect judgments in a variety of fields, such as social judgments (e.g., of persons, Higgins et al., 1977), psychophysical judgments (e.g., of weights, Helson, 1964), psychosocial judgments (e.g., physical attractiveness, Kernis & Wheeler, 1981) or judgments of self-relevant concepts (e.g., subjective well-being, Schwarz & Strack, 1999). Apart from finding evidence for assimilation and contrast effects in evaluative judgments (cf. chapter two), the focus in this chapter was on the underlying and mediating processes. A number of different models has been developed over time to explain the causes of, boundary conditions for and underlying processes of assimilation and contrast effects (see for example Ford & Thompson, 2000). The concept of social comparison is similarly prominent research field. Starting with the work of Festinger (1954), there seems to be an ever-growing interest in the field of social comparison (see Buunk & Mussweiler, 2001). There is, of course, an apparent link between these two fields. The outcome of social comparison, i.e., the judgment of the self is of course – as any judgment – affected by assimilation and contrast effects (e.g., Kühnen & Haberstroh, 2004; Stapel & Schwinghammer, 2004; Suls, Martin, & Wheeler, 2002).

The SAM (Mussweiler, 2003a) offers a link between these two related research topics. Focusing on the basic underlying psychological mechanisms, the model tries to offer a social cognitive perspective of assimilation and contrast effects in judgments as a result of social comparison. It is important to note that the assumptions of this model apply not only to *social comparisons*, but to any kind of (social) judgments (Mussweiler, 2003b). The results of the last chapter already show that this might be the case for performance judgment of students in an exam. There is also some recent work showing the applicability of the model for judges in gymnastics (Damisch, 2004). The SAM assumes that each judgment is preceded by a comparison between target and standard. The following judgment of the target will then be based on the information activated during this comparison. Since people have a certain expectation or hypothesis in their mind guiding the information search (in the direction of a positive – but not necessarily confirmatory – test-strategy), the activation and interpretation of information is likely to be biased in a direction consistent with the expectation. This is the cause for biases, i.e., assimilation and contrast effects, in the judgments of a target. The model states that a similarity-

hypothesis will lead to assimilation effects and a dissimilarity-hypothesis will lead to contrast effects. Although there is a growing number of evidence for the SAM (see Mussweiler, 2003a), there seems to be little work focusing on the assumed process of hypothesis-testing and information activation. Using a lexical-decision task (Dijksterhuis et al., 1998), it has been shown that hypothesis-consistent information is activated after judging for example the price of a car (Mussweiler & Strack, 2000c), one's athletic ability (Mussweiler & Strack, 2000b), or one's neatness (Mussweiler & Bodenhausen, 2002). However, these results do not prove that the concepts have been activated prior to the judgment. The activation of the hypothesis-consistent concepts might have been the consequence rather than the cause of the biases in the judgments. There is also one study using a thought-listing procedure to show that more hypothesis-consistent than hypothesis-inconsistent thoughts come up while judging the length of the river 'Elbe' in an anchoring-paradigm (Mussweiler & Strack, 1999). Still, one has to consider the disadvantages of this method, namely that the thought-listing procedure may change the judgment task itself. The present work is, therefore, the first direct test of this process of hypothesis testing assumingly underlying assimilation and contrast effects, by using a rather unobtrusive method.

In the studies four to six reported in this chapter, participants were in the role of a teacher, testing the knowledge of virtual students in an interactive computer simulation. In this simulation participants had to ask questions to and received answers from the virtual students. The focus of these studies was on the difficulty of the questions chosen by participants, because this was expected to reflect the process of hypothesis testing during an exam. Participants were expected to test their own expectations (regarding the target student) by changing the level of difficulty of the questions asked. Expectations of participants regarding the performance of the target students were manipulated via two independent variables, namely the performance of the first virtual student (high vs. low standard) and the focus of participants (on similarities vs. on dissimilarities). When expecting a student to perform badly, it was predicted that participants would test this by asking more difficult questions; when expecting a student to perform well, it was predicted that participants would test this by asking more easy questions. Thus, participants were expected to employ a positive – but not necessarily a confirmatory – test-strategy.

The results of the studies four to six confirm this assumption. Participants followed a test-strategy that is consistent with their expectations. Yet, compared to classical studies on hypothesis-testing (e.g. Snyder & Swann, 1978), using this strategy did not activate hypothesis-consistent information, thus, verifying the expectations. In the studies four to six, participants expected the performance of the

targets students to be either good or bad. However, in the paradigms used, the performance of the target students was either intermediate (i.e., neither good nor bad) or dependent on the questions asked in such a way that more difficult questions would make incorrect answers more likely (and vice versa). Therefore, the information received while following a positive test-strategy did in neither case fully support the expectations participants held. How did participants react to this non-confirming information? Theoretically they could have tried to integrate the given information by interpreting it in a hypothesis-confirming way. Yet, this could not easily be accomplished with the information in the paradigms used because with each answer participants were explicitly told if the answer was correct or incorrect. What participants instead did, was an adaptation of their expectancies to the relevant information they received, i.e., they altered their expectations to fit reality.

Two findings in the studies four to six support this view. On the one hand, the test-strategy employed by participants was reflected mostly in the first half of an exam when they had received only few pieces of information. During the cause of the exam their initial expectations influenced their test-strategy less and less. On the other hand, the bias in the performance judgments, i.e., the expected assimilation and contrast effects, were rather weak, though they were overall in the expected direction. Both findings suggest that participants adjusted their expectations to the constraints given by reality, i.e., they were not entirely blinded by their expectations (maybe only a little bit short-sighted). This is a rather interesting finding because most studies on expectancy effects or hypothesis testing do not focus on the impact of disconfirming information on the expectancies. Of course, this was not directly tested in the studies four to six either, but the results suggest that participants were able to change their expectations. Even so the expectations changed over time, the results of the mediator-analyses suggest that the test-strategy employed at the beginning of each exam is a mediating variable between the expectations held and the biases in the final judgments of the target students.

The findings of study four – or rather the non-findings – lead to another important aspect of the SAM. Even so the model explicitly states that every judgment is preceded by a comparison of a target with a standard, it does not specify what standard a judge uses. Mussweiler (2003a) refers to the literature and states three important mechanisms for the selection of a standard: Conversational inferences (e.g., Schwarz, 1994), accessibility in memory (e.g., Herr, 1986) and normative concerns, including that standard and target are similar enough to be comparable (e.g., Smith & Zárate, 1992).

The nil-findings in the fourth study were explained by the assumption that participants did not automatically use the first virtual student as a comparison standard. Even so the first student was described as being similar to the target and was very likely accessible in memory, participants did not use her as a standard. It may be that normative concerns prevented them from doing so, because a fairness norm may exist that interdicts to compare the performance of different students in an exam with each other. As mentioned above, the performance judgment in an exam should be as objective as possible. There was no information in the instructions telling participants not to use this fairness norm and to compare the virtual students. By additionally telling participants that they would have to propose one of the students for an award in the fifth study or that they would have to rank order the performance of participants in the sixth study, they could infer that it was demanded by the experiment to compare the students and this would, therefore, be no violation of the norm.

Still, the SAM assumes that there is always a standard the target is compared with. There was no alternative standard available in the studies (except the performance of the first student). So, what kind of standard did participants use? The recent work of Mussweiler and Rüter (2003) suggests that people develop routine standards, like the self or close friends that they use for many judgment tasks. If participants chose an individual standard in the fourth study, this would have led to random effects when combined with the manipulation of the focus and would explain the nil-findings. Still, it might also be possible that participants tried to correct for the influences of a comparison process, because of an activated fairness norm as suggested in correction models such as the *Flexible Correction Model* (Wegener & Petty, 1997).

Building up performance expectations in participants did not only influence the performance of one, but of two target students following the standard student as could be seen in the sixth study. This seems to contradict the findings of the last chapter, namely that the effect of the manipulated expectancy was not very stable and wore off quickly over time. However, this is only true if these expectations are not used or tested. By being put in a more active role using the computer simulation, participants were most likely testing their assumptions over the whole course of the experiments, thus, keeping the expectancies accessible and letting them influence their test-strategies and judgments. Participants even used these expectancies when they were asked to judge their own abilities. Here, the effects in the judgments were even stronger than in the performance judgments of the virtual students. This can be explained by the fact that in the case of the ability judgments of the self, participants

did not get contradicting information from the environment. Instead, their whole information search was memory-based, making their expectations likely to affect the activation of (expectancy-consistent) information as well as the (expectancy-consistent) interpretation of this information. This may explain the stronger effects found in this case.

To conclude, studies four to six shed light into the processes underlying assimilation and contrast effects in evaluative judgments. The findings support the core assumptions of the SAM, namely that there is a comparison of target and standard preceding a judgment and that the information search during this comparison is guided by a hypothesis. Yet, the positive strategies used to test the hypotheses are not necessarily confirmatory and do, therefore, not necessarily lead to biased judgments. If the information received contradicts the expectations of the judge and this information cannot easily be interpreted in line with the expectations, people adjust their prior view to reality.

Chapter 4

-

Concluding thoughts

4.1 Systematic influences of prior judgments on subsequent ones

The main question guiding my work presented here was whether and in what direction prior judgments would systematically influence subsequent judgments in a sequential judgment situation. The main idea was that prior judgments should function as a comparison standard with which the following targets to be judged would be compared. This comparison should be a possible cause for a judgmental bias in the performance judgments of the target to be judged, either in an assimilative or a contrastive way. Prior work on sequential judgments in complex, applied situations has already shown evidence for judgmental biases caused by the sequential character of the task. This research comes from different areas where evaluative judgments are required, such as judgments in sports (e.g., Plessner, 1997), the judgment of exams (e.g., Betz, 1974), judgments of job performance (e.g., Sumer & Knight, 1996), and personnel decision making (e.g., Highhouse & Gallo, 1997). The research on the effects of sequential performance judgments has shown that prior judgments may have an assimilative (e.g., K. R. Murphy et al., 1986b), a contrastive (e.g., K. R. Murphy et al., 1985) or both effects (e.g., Smither et al., 1988) on the following ones.

Although many different variables (e.g., presentation form of performance, time lag between judgments, extremity of performance) seem to have an important influence on the direction of the effect, there seems to be no general rule or heuristic making it possible to predict when to expect assimilation and when to expect contrast effects in sequential judgments, at least when it comes to complex, applied judgment situations. Therefore, Murphy and colleagues (1985) came up with the following statement: “Although we expect that systematic differences in previous performance will affect evaluations of present performance, it is impossible, …, to confidently predict either assimilation or contrast effects” (pp. 76-77). In my opinion, this is a very unsatisfactory statement that seems to be representative to this field. Since there

seems to be no easy solution at hand, the same authors further stated that “almost regardless of whether one finds a contrast effect or an assimilation effect, the mere fact that previous performance has a systematic effect on evaluations of present performance is both theoretically and practically important” (K. R. Murphy et al., 1985, p.82). I agree with this statement, but it does not really help to solve the puzzle of assimilation and contrast effects.

Moving away from the applied field, there has been quite an amount of work on assimilation and contrast effects in the fundamental disciplines, i.e., in the fields of cognitive (see Petzhold & Haubensak, 2001) and social psychology (see Ford & Thompson, 2000; Stapel & Koomen, 2001b). Over the years, this has led to the development of manifold different theoretical models and approaches. These models not only try to give an explanation on *how* assimilation and contrast effects evolve in judgments, they also make predictions on *when* to expect assimilation and contrast effects, respectively. These models seem, therefore, useful to answer the same questions for more complex, applied judgment situations. Unfortunately, many of the models are built on empirical evidence that is derived from studies using rather simple judgment tasks. Therefore, it seems to be an important but unanswered question whether the assumptions of these models can also be applied to more complex judgment situations.

Although it has been said that there is nothing as practical as a good theory, it seems to me that one important aspect of psychological theories is often disregarded, namely if a theory can be applied to complex tasks, a central feature of applied judgment situation. Therefore, the first main goal of my work was to use a recent theoretical framework (the Selective Accessibility Model [SAM], Mussweiler, 2003a) that allows predictions about when to expect assimilation and contrast effects respectively, and to apply the assumptions of this model to a complex judgment task, namely the grading of exams. The SAM was chosen because it is a recent model that incorporates assumptions of earlier models, e.g., the *Inclusion/Exclusion Model* (Schwarz & Bless, 1992a). It also sees comparisons as the process mediating influences of prior judgments on subsequent ones in sequential judgment situations. If the SAM (or related models) proofs to be applicable to more complex judgment situations, this would help answering the question of when to expect assimilation and contrast effects in applied situations.

The first three studies (reported in chapter 2) were designed to answer the questions aligned with the first goal outlined above. In these studies, participants were put in the role of a teacher and given the task to grade the exam performance of two pupils. Two variables were manipulated: On the one hand, the performance of the first pupil (comparison standard) was either good (high standard) or bad (low standard); on the other hand, the focus of participants was either set on similarities or on dissimilarities by using a ostensibly unrelated priming task. It was expected, in accordance with the SAM, that participants would compare the second pupil (target) to be judged with the first one (standard). This should lead to assimilation effects when participants were focusing on similarities and to contrast effects when they were focusing on dissimilarities.

This expected interaction pattern was found in the first study. In the second study only assimilation effects appeared and the third study showed no biases in the judgments of the target. Taken together, the first three studies give only partial or weak evidence for the expected pattern that was derived from the assumptions of the SAM. In general, it seems that assimilation effects were more likely and stronger than contrast effects in the judgments of the target. The results were unexpected on the basis of the SAM. This does not necessarily mean that the model cannot explain them. Although the results were not as expected, the deviations can be explained if the reliability of the paradigm and the manipulation of the comparison focus used is questioned.

Overall, it seems that the strength of the effects of the priming manipulation depended on the time lag between the manipulation itself and the judgments to be made: The longer the time between the manipulation and a judgment, the weaker the expected effect. This can explain why results in study one show a strong (though unexpected) interaction pattern of assimilation and contrast effects for the judgment of the standard (first pupil) and weaker effects for the judgment of the target (second pupil). This can also explain why the manipulation did not affect the judgments in the second study. This time, audio-recorded protocols of the exams instead of written ones were used. The time needed to listen to these was longer than the time to read through the written protocols. The most likely explanation for the nil results in the third study is that participants were aware of being manipulated and were able to de-bias their judgments (cf. Wegener & Petty, 1997).

Even if the manipulation of the focus – and, therefore, of the direction of the judgmental bias – did not work, an assimilation effect was found over the first three studies. Assimilation effects in judgments are expected to be the default effect by the SAM, because a certain amount of similarity is necessary for a comparison between a target and a standard to take place. Two other aspects of the paradigms used in the first three studies might explain why the effects were not as strong as expected. The SAM sees the hypothesis-guided search for and activation of information as the main cause for assimilation and contrast effects in the later judgments. In the paradigms used in the first three studies, all information about the performance of the two pupils was available to all participants, i.e., participants did not actively have to search for information. Therefore, no bias due to the selective activation of information could be expected. In most studies testing the assumptions of the SAM, there was no information given about the target to be judged. Especially in social comparisons paradigms (i.e., judgments of the self), information has to be sought by a memory-based search. The only source of a bias in my paradigm was at the interpretation stage of the information given to participants. Although the answers of the target student in the protocols were designed to be ambiguous, it may be that the room for interpretation was too small. It may also be possible that interpretation of information is a much weaker source of bias than the selective search and activation of information. This assumption is supported by the work of Williams and colleagues (Williams et al., 1990) showing that on-line judgments are more accurate than memory-based judgments.

Taken together, the results do not really contradict or question the assumptions of the SAM. If one considers the weaknesses of the paradigm used in my first three studies as outlined above, the SAM still can account for the results. However, even if the SAM is able to explain these results, the first three studies still show a weakness of the model. Even though the model makes predictions about when to expect assimilation and contrast effects, respectively, these predictions are not very precise. According to the model, the decision to test or focus on either a similarity- or dissimilarity-hypothesis is made during a first quick holistic assessment of the similarity of target and standard. A similarity judgment during this stage leads to similarity testing and assimilation effects, a dissimilarity judgment to dissimilarity testing and contrast effects. However, the model does not answer two open questions: First, how much similarity (or dissimilarity) between target and standard is

needed to build a similarity- (dissimilarity-) hypothesis; second, how (or based on what features) do participants judge the similarity between target and standard? Answers to these two questions are a necessary perquisite to make precise predictions on the direction and size of the judgmental bias. It may, therefore, be an important step to incorporate work on similarity-judgments into the SAM, such as the work on feature-matching models by Tversky and Gati (e.g., Gati & Tversky, 1987; Tversky, 1977; Tversky & Gati, 1978, 1982) or by Houston and colleagues (e.g. Houston et al., 1989; Houston, Sherman, & Baker, 1991) on features of similarity.

To conclude, the SAM seems to be a powerful model to explain assimilation and contrast effects in judgments (even for complex judgment tasks), but not as powerful in making precise predictions for the direction and size of these biasing effects, at least for the paradigms used in my first three studies. It seems, therefore, doubtful whether this model is able to resolve the puzzle given by the mixed findings on assimilation and contrast effects in applied judgment situations.

Following this critique, one may ask whether other models allow more precise predications about assimilation and contrast effects. I will consider two of the models that have been outlined in the first chapter[71]: The *Inclusion/Exclusion Model* (Schwarz & Bless, 1992a) and the *Interpretation/Comparison Model* (Stapel & Koomen, 2001a). The *Inclusion/Exclusion Model* (Schwarz & Bless, 1992a) states that the process of categorization of target and standard decides whether assimilative or contrastive biases will be more likely. If target and standard are seen to belong the same category, assimilation effects are more likely; if target and standard are seen to belong to different categories, contrast effects are more likely. The model does not give any general explanations on how this important process of categorization works. Instead, the authors offer a variety of variables that may determine the inclusion or exclusion of target to the category of the standard (e.g., temporal distance, category width, presentation, and judgment order). Again, it seems, therefore, necessary to incorporate theoretical work on categorization processes[72] (such as the work by Tversky & Gati, 1978) into the given framework to be able to make precise predictions on direction and size of the judgmental bias in a given situation.

[71] The third model (The Flexible Correction Model by Wegener & Petty, 1997) builds on existing models like the SAM, adding a stage of corrections. Therefore, I would expect that this model would deal with the same and even additional problems when it comes to making predictions.
[72] Categorization and similarity seem to be rather related topics.

The *Interpretation/Comparison Model* (Stapel & Koomen, 2001a) states that the same kind of activated context information may lead to assimilation and contrast effects, depending on during which stage or process of a judgment it is used. If information about the standard is used during the interpretation stage, it will have assimilative consequences; if the same information is used during the judgment stage, it will have contrastive consequences. Again, instead of giving an explanation of how this differential use of activated information may work and how it is decided when it will be used, the authors offer a number of variables that may determine during which stage context information may be used (e.g., extremity, abstractness of the information, or goals of the judges). The model also states that similarity may play an important role: If standard and target are seen as being similar, standard information will be used a comparison standard during the judgment stage and, thus, lead to contrast effects; if, however, standard and target are seen as being dissimilar, standard information will be used to interpret the information about the target and lead to assimilation effect. Although this is an interesting assumption contradicting the *Inclusion/Exclusion Model* and the SAM, the same critique as for the two other models may be applied to this model as well.

All three models seem to be able to *explain* a broad number of findings concerning assimilation and contrast effects, including the results of my first three studies. Thus, they provide, as Schwarz and Bless (1992a) put it, a “heuristically fruitful integrative framework for the conceptualization of assimilation and contrast effects” (p. 218). Still, in my opinion, these models do not yet offer the possibility to make precise *predictions* about when to expect assimilation and contrast effects in judgments, thus, questioning their usefulness for the demands of complex, applied domains. What needs, then, to be done to answer the question – coming from applied judgment situations – regarding the direction and size of judgmental biases?

First, I believe that the theoretical models – at least the ones presented here – need further development, especially on their core parts where it is decided, how and when assimilation and contrast effects will develop (e.g., the ‘quick holistic assessment’ in the SAM). Second, I believe that these models, coming from the field of fundamental research, should be more often applied to or tested using complex, applied judgment situations, thus, making strengths and weaknesses more obvious. Third, a different solution would be to focus on the most important variables in applied judgment situations that influence direction and size of judgmental biases

and to find better ways to define, quantify or operationalize these. To conclude, even if there are already many studies and well developed models on assimilation and contrast effects in judgment, more work needs to be done to allow precise predictions.

4.2 Processes underlying assimilation and contrast effects in judgments

The first goal of my work was to see whether comparisons might influence evaluative performance judgments when these judgments are made in a sequence. In the first part of my work I could show that previous judgments have an (assimilative or contrastive) influence on subsequent judgments. These results are in line with recent theoretical approaches on assimilation and contrast effects, such as the SAM (Mussweiler, 2003a) or the *Inclusion / Exclusion Model* (Schwarz & Bless, 1992a). Both models predict that assimilation effects are likely when the target to be judged and the comparison standard[73] are considered to be similar[74]; contrast effects are likely when target and standard are considered to be dissimilar. Following these models, prior judgments should have an assimilative effect on subsequent ones when the judged persons (or objects) are seen as being similar; prior judgments should have a contrastive effect on subsequent ones when the judged persons are seen as being dissimilar. Yet, why is this the case? What are the underlying or mediating processes that lead to assimilation or contrast effects in judgments? To answer this question was the second goal of my work and I addressed this question in the second part of my work (Chapter 3).

The assumptions concerning the underlying processes were derived from the SAM. I used this model for several reasons. First, it is a recent and well developed model that already includes parts of earlier models on assimilation and contrast effects. Second, building on the idea that comparisons have a strong influence on evaluative judgments, the model makes precise assumptions about the mediating processes underlying assimilation and contrast effects and comparative judgments, respectively. In the situation of sequential performance judgments, I believe that the comparison of prior judgments with subsequent ones is one especially important process underlying judgmental biases. Third, the model focuses especially on the effects of selective information search or activation. In the situation I focused on with my work – the case of grading oral exams – a judge is not passively receiving information, but actively searching for information. Therefore, I believe that the SAM is a model well suited to explain evaluative judgments in this situation.

[73] Or the context information, in the *Inclusion / Exclusion Model*.
[74] Or to belong to the same category, in the *Inclusion / Exclusion Model*.

In the paradigm used for this part of my work, participants were again in the role of a teacher confronted with the task to grade exams of two (or three) students. However, instead of being in the passive role of judging performances based on written or oral protocols, this time participants had to lead actively through the examination by asking questions and getting feedback from the tested students. To achieve this, a computer-based simulation of an oral examination was designed and used in three studies (studies four to six). Participants had to make evaluative performance judgments of two (or three) students in a sequence. It was expected that participants would build an expectation based on the performance judgment of the first student (standard) and that this expectation would influence their test strategy (i.e., the difficulty of the questions asked during the exam) for and the performance judgments of the following students. The SAM assumes that judges always follow a hypothesis when comparing the target to be judged with a given standard. This hypothesis will guide the process of information acquisition; therefore, this information will be activated selectively in the sense of a certain (positive) test-strategy (Klayman & Ha, 1987). This will lead, in many cases, to the activation of hypothesis-consistent information. Since the final judgment of the target will be based on this activated information, a hypothesis-consistent judgment will also be likely.

In studies four to six, I manipulated the expectations (or hypotheses) of participants. They were led to expect either a good or a bad performance of the following target to be judged. These expectations were manipulated via two variables, namely the performance of the first student (high vs. low standard) and the comparative focus (on similarities vs. on dissimilarities). When expecting a good performance, I predicted participants to test this by asking more difficult questions than when expecting a bad performance. This could be called a positive test strategy because participants would be “examining instances in which the property or event is expected to occur (to see if it does occur) …” (Klayman & Ha, 1987, p. 212). The underlying assumption is that giving a correct answer to a difficult question is an indicator for a good student and an incorrect answer to an easy question is an indicator for a bad student. I also predicted that the expectation of a good performance would lead to better performance ratings than the expectation of a bad performance.

The results show that participants do, indeed, adjust their test strategies to their expectations in the expected direction. When participants expected a student to

perform well, they asked more difficult questions than when expecting a student to perform badly. This may have at least two reasons. First, asking difficult (easy) questions to test the hypothesis that a student is good (bad) is more informative or diagnostic than asking easy (difficult) questions. Imagine a judge following the expectation (testing the hypothesis) that a student is good. If that student answers a difficult question correctly or incorrectly, this will give the judge confirming or disconfirming information regarding the hypothesis. However, if the student answers to an easy question, this will only be informative for the judge if the answer is incorrect. Answering an easy question correctly is necessary but not sufficient to judge a student to be good. The same logic – the other way around – applies to a judge following the expectation that a student will show a bad performance.

Thus, participants were using a positive test strategy, but this strategy was also a very diagnostic strategy for the paradigm used (cf. Fiedler et al., 1999, p. 6). Therefore, this finding is in line with studies showing that "the diagnostic value of behavior is a stronger determinant of question selection than the probability of the behavior under the hypothesis" (Trope & Liberman, 1996, p. 246). A second explanation for the employed test-strategy might be that participants simply associate difficult questions with good performance and easy questions with bad performance. In this case, participants would be following a kind of "matching-heurism" (Evans, 1989; Evans & Lynch, 1973).

A second important finding is the fact that these test-strategies, which manifest in the differential use of easy and difficult questions, do change over time, if disconfirming evidence is found. The computer-based simulation used in studies four to six was build in a way that disconfirming evidence was more likely than confirming evidence, independent of the expectations held by the judges. Thus, it is important to note that participants did not employ a *confirmatory strategy* (Snyder & Swann, 1978; Wason, 1960) by following a *positive test strategy* (see Oswald, 1993, for an overview). On the one hand, the probability to answer an easy question correctly was always higher for the virtual targets than the probability to answer a difficult question correctly[75]. Therefore, participants who expected a well performing student asked more difficult questions and had a higher probability to receive an incorrect answer than participants expecting a badly performing student. Participants were, therefore, following a *hypothesis-disconfirming test-strategy*. On the other hand, the

[75] In studies four and five.

performance of the targets was neither good nor bad (especially to the explicitly good or bad standard), but only average. Therefore, the number of correct and incorrect answers was always disconfirming information, assuming that participants expected either a good or a bad performance.

So, participants started with a performance expectation and were collecting more and more disconfirming information. This had an influence on the test-strategy participants used. Over the course of time, participants adjusted their test-strategy (and assumingly their performance expectations) to match the gathered information. If the performance of a student was better than expected, participants started asking more and more difficult questions; if the performance was worse than expected, participants started asking more and more easy questions. Therefore, the differences in the test-strategies caused by the different expectations were getting smaller over time, resulting in similar test-strategies for all participants in the end. This could also explain why the differences in the performance judgments between participants with different expectations were rather weak or non-significant.

To conclude, judges may have certain expectations that influence their test-strategies, but judges are not blind to reality (maybe only a little short-sighted). They are able to adjust their expectations and judgments to (disconfirming) information from the environment[76]. This is an important finding that is rarely reported in the literature concerned with expectancy effects or hypothesis testing. Here, mostly expectancy- or hypothesis-congruent judgments are reported or focused on. Yet, there are some examples in the literature that show a change of the hypothesis being tested because of disconfirming feedback. Klayman and Ha (1987) speak of a possible change of an initial strategy in studies of concept identification (via feedback) where participants follow a “win-stay, lose-shift strategy” (p. 221).

There is also a noteworthy study by Gadenne and Oswald (1986) focusing explicitly on the reaction of hypothesis-testers to hypothesis-disconfirming information. Participants were in the role of a detective trying to solve a case of murder. They first had to read through part of a criminal story to build a hypothesis about who the primary suspect might be. Later on, participants were receiving additional information that was more or less convincing, but always disconfirming

[76] An alternative explanation for this ‘adaptive test-strategy’ may be that participants had a motivation to ask as many difficult as easy questions. After having started with some difficult (or easy) questions because of their test-strategy, they might have adjusted their later questions accordingly to achieve an equal distribution of the question difficulty (cf. Parducci, 1965). This alternative account may be tested by an additional manipulation of the target performance (to be either confirming or disconfirming).

regarding the initial hypothesis. Gadenne and Oswald found that participants were not ignoring important disconfirming evidence; they also found a tendency for participants to adjust or change their initial hypothesis to the disconfirming information, especially when this information was important and an alternative hypothesis was ready at hand.

Comparable results can be found in the work by Fiedler and colleagues (Fiedler et al., 1999, Experiment 1) on the auto-verification effect in social hypothesis testing. In their first study, participants had to test the hypotheses that men show an overt and women a covert aggression style. Participants had to collect information about two persons, “Peter” and “Heike”. Among other things, the affirmation rate of the feedback (regarding the hypothesis) was manipulated to be either high or low. The collection of information was guided by a positive test-strategy: Participants looked more at instances of overt aggression for Peter than for Heike, and the other way around for covert aggression. This resulted in an auto-verification effect in the frequency estimates given in the end: Participants judged the (relative) frequency of overt vs. covert behavior to be higher for Peter than for Heike. However, this effect was more pronounced when the affirmation rate of the given feedback was high than when it was low. In fact, when the affirmation rate was low, there were no significant differences in the frequency estimates. In contrast to my findings, the affirmation rate had no effect on the search-strategy employed by participants.

Taken together, these findings do correspond to the results I found in my studies four to six, giving additional evidence for the conclusion that people may be *positive* (Klayman & Ha, 1987) but not necessarily *confirmatory* hypothesis-testers (Snyder & Swann, 1978; Wason, 1960), but are open to important disconfirming information.

The described positive test-strategy was only found in two of the three studies reported here (study five and six). A difference in the paradigms that may account for this might be that participants were implicitly instructed in these two studies to compare target and standard, by adding an additional task. This task was either to suggest one of the students for an award or to rank-order the students in the end of the study. In the fourth study, there were no such instructions leading participants to compare the students with each other. How did participants then come up with an evaluative performance judgment?

According to recent work by Mussweiler and Rüter (2003), judges may use routine-standards – i.e., standards that they have often used before – as a default when no other standard is offered by a task. In the case of self-judgments, this might be the best friend, in the case of judgments of other persons it might be the self (Dunning & Hayes, 1996). When judging the performance of students during an exam, it may be likely that participants in my studies were using their own assumed performance as a comparison standard, because participants were students as well. This idea is fostered by some unexpected effects of the manipulation of the focus in study five – where targets to be judged were students – that was not replicated in study six – where targets to be judged were pupils. Yet, participants may also have used any other available and applicable comparison standard, e.g., some abstract prototype of the typical student or pupil.

Although the SAM is able to include individual routine standards to explain comparative judgments without a salient or obvious standard, this assumption again weakens the predictive power of the model, at least in complex judgment settings. To predict the direction and the size of the influence of a standard on the judgment of the target, one needs to know which standard is used. Since this may differ from person to person and from situation to situation, a precise prediction of the judgments seems to be rather difficult on basis of the SAM. It is important to note that this critique does not only concern the SAM, but also the other models illustrated here, such as the *Inclusion/Exclusion Model* (Schwarz & Bless, 1992a) and the *Interpretation/Comparison Model* (Stapel & Koomen, 2001a). Additionally, it is important to note that the findings of studies four to six do not apply to the *Inclusion/Exclusion Mode*l and the *Interpretation/Comparison Model*. Both models are concerned with the differential use of accessible information as a cause for assimilation and contrast effects. The findings of studies four to six show a selective search for and activation of information as suggested by the SAM, the only of these models that makes assumptions about the differential *activation* of information, rather than the differential *use* of already activated information. This does not mean that these models are mutually exclusive. The assumptions of these models should rather be seen as supplementary explanations of assimilation and contrast effects in judgments.

4.3 Deductions and an outlook: Some suggestions for improving evaluative performance judgments

What are the implications from the results presented here? What can be deducted to make judgments in applied settings less prone to biases that are based on comparisons? First of all, it is important to note that comparative judgments are not always unwanted and unintentional in situations where sequential judgments are made. When searching for *one* candidate for a job, or when searching for *one* candidate for an award, judges are concerned with a rank order: The goal is to find the *best* candidate among the applicants. Still, even in these situations, all candidates should be compared with the same standard; otherwise different criteria would be used to judge different candidates. In the domain of judging exams, the goal is not to rank order the students tested, but to grade the performance as objectively as possible. This means that the performance judgments should be based solely on the relevant performance behaviors (i.e., the answers given by the student) and not be influenced by context variables (e.g., prior judgments in the sequence) or by the judge. Only in this case the given grade is informative per se and can be used to compare one student with students from other schools and age groups; and only then certain grades of a normative grading scale can be assigned a certain meaning (see also Birkel, 1984b).

The results of my studies clearly show that judges do compare the performance of students with each other when judging their performances in a sequence. Especially the student judged prior to a target student is often used as a comparison standard. These comparisons lead to unwanted assimilation and contrast effects in the performance judgments of the target student. Therefore, these performance judgments cannot be called objective. Even bringing students of one sequence into a rank order would not be possible, because different comparison standards are used when judging different students.

Taking the notion of Mussweiler (2003a) for granted that all judgments are relative (i.e., based on comparisons), how can objective judgments then be achieved? One solution would be that judges should choose the same (abstract) standard to compare students taking an exam with. It would be useful to define a kind of prototype of a student taking an exam based on the required knowledge and skills (which have prior to be defined). If all students were to be compared with the same

standard, the amount of the judgmental bias would be the same for all students and could, therefore, be neglected. Still, the focus of judges (on similarities or on dissimilarities) would influence the direction of the bias. Here it seems important to find the relevant variables that have an influence on the focus of comparison.

A mediating variable for the observed assimilation and contrast effects seems to be the quality of the questions asked during an exam, especially the difficulty of the questions, which directly influences the performance of the tested students (i.e., the probability of correct / incorrect answers). This means that each student is confronted with a different test situation when taking an exam (cf. Birkel, 1984a), making comparisons of the performances impossible when the judgments in oral exams are based on the logic of the *classical test theory* (Gulliksen, 1950). However, the exam situation could also be seen as an adaptive test, following the logic of the *item response theory* (see Fischer, 1996). In this case the latent performance parameter of each student could still be estimated. To use the logic of the *item response theory*, the difficulty of questions needs to be determined as an important parameter. Yet, to my knowledge, there exists no study trying to determine the difficulty of questions used in oral exams based on the number of students that are able to answer these questions correctly.

Additionally, my own data shows that there are huge individual differences in the difficulty judgments of the questions used. It seems, therefore, unlikely that teachers are able to intuitively judge the difficulty of questions in a reliable way (cf. Schnotz, 1971) and that they use this difficulty to base their judgments on following the logic of item response theory. A practical and easy solution would be to use the same or a similar set of questions for all students being tested to make the observed performances comparable.

4.4 Critical remarks: The work is not complete

The ideas for my work were derived from a complex, applied judgment situation, namely the grading of (oral) exams. Likewise, a goal underlying my work was to close the gap between fundamental and applied research a bit by using the assumptions from fundamental theories on human judgments as a basis to explain judgments and judgmental biases in more complex tasks. Complexity is one of the central defining features of applied judgment situations. Since it has been criticized whether findings from laboratory studies can be generalized to applied settings (cf. Ilgen & Favero, 1985; K. R. Murphy & Cleveland, 1991), it is, therefore, important to think about the ecological validity of my findings.

First, participants who were making the performance judgments in my studies were students who should be used to *take* but not to *grade* exams. Can the findings, therefore, be generalized to *expert* teachers at schools and universities? Teachers can be called experts in making performance judgments if they haven been making many of these judgments before. Therefore, they may have developed a certain strategy of what questions to ask during an exam that may differ from a laypersons' strategy. I would expect that expert teachers would have developed a routine of doing an exam that would determine at least part of the (difficulty of the) questions asked. Therefore, I would expect the effects on the difficulty of questions found in studies four to six to be weaker for professional teachers. I would also expect expert teachers to use a different strategy than laypersons when judging the performance.

Teachers cannot be called experts for making performance judgments in a sense that they are better (i.e. more objective) at this task than laypersons. To increase the objectivity or accuracy of a judgment, a judge has to be able to learn from his behavior, i.e., he has to receive feedback for his own judgment performance. This is not the case for teacher grading exams. In a study investigating the grading behavior of 164 expert teachers, Birkel (1978) found nearly no correlation ($r = .07$) between the average judgments given by the teachers. Therefore, I do not think that expert teachers are aware of and prone to the influence of comparisons. Yet, it may be that they use different comparison standards than laypersons, e.g., a prototype of a typical student or prior performance information of the same student.

Second, the exam situation used in my studies differed from a real exam situation in some ways. Although the computer-based simulation used in studies four to six made it possible for participants to actively interact with the student being tested, there were obvious differences to a real exam situation: In a real exam situation the number and kind of questions asked is not restrained, the student taking the exam is usually quite nervous, and there are many additional variables that may influence the judging teacher (e.g., nonverbal cues given by the student being tested). A real exam situation is therefore even more complex and these additional variables may have a moderating influence on the effects observed in my studies, either strengthening or weakening them.

Taken together, although I do believe that comparisons play an important role even in real applied judgment situations, this question cannot be finally answered by the work presented here. The question whether the effects of my studies can be generalized to expert judges and to real exam situations can only be addressed by further empirical work.

4.5 Conclusion

To conclude, over the course of six studies I could show that judges do compare the performance of students when making sequential performance judgments. It seems that the student prior to the target student functions as a comparison standard for the target, therefore, influencing the judgment of the target. These comparisons can lead to assimilation or contrast effects in the judgments, a bias that is unwanted for most applied judgment situations (e.g., grading of exams). To answer the question of the direction of these comparison effects, i.e., of when to expect assimilation or contrast effects, seems not to be an easy task. Although there are well-developed theories coming from fundamental research that deal with this problem, a closer look reveals that the predictive power of these theories seems to be weak when it comes to complex, applied judgment situations: These theories may have shown their predictive power by using simple, well-developed paradigms; they are also able to explain (post hoc) almost all results found in my (and other applied) studies. It seems, therefore, questionable whether some assumptions of these models can be falsified at all. Finally, the predictive power of these models seems to be weakened when it comes to new judgment situations.

Nevertheless, it is important to understand comparison processes if one wants to understand the process of judgments. In studies four to six this comparison process was the focus of attention. Based on the assumption of the SAM, it was tested whether judges follow and test a hypothesis or expectancy when comparing a target to be judged with a given standard. It was, indeed, found that judges chose the questions they ask in a strategic way, following a *positive* test-strategy. Furthermore, it could be shown that judges were not necessarily following a *confirmatory* test-strategy, but were able to adjust their initial hypotheses or expectancies to the feedback they acquired from the environment. Therefore, even if the SAM may not be able to make precise predictions about the final judgment in complex judgment situations, the assumed underlying mechanisms can still be shown in these situations.

Overall, my studies show that it is important and fruitful to test well-developed theories in different laboratories (or in the field), using different paradigms (or complex, applied situations), thus, making the further development – instead of the

recurrent confirmation – of established theories possible. Using the words of Klayman and Ha (1987), I would say that we may use a positive test-strategy to test our theories, but only if this test-strategy is also a diagnostic strategy. Mere confirmation is hindering the scientific progress.

References

Aarts, H., & Dijksterhuis, A. (2002). Category activation effects in judgment and behaviour: The moderating role of perceived comparability. *British Journal of Social Psychology, 41*(1), 123-138.

Allport, G. W. (1937). *Personality: A Psychological Interpretation.*: Henry Holt.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal & Social Psychology, 41*, 258-290.

Bargh, J. A., & Chartrand, T. L. (2000). The mind in the middle: A practical guide to priming and automaticity research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253-285). New York, NY: Cambridge University Press.

Bazerman, M. H., Beekun, R. I., & Schoorman, F. D. (1982). Performance evaluation in a dynamic context: A laboratory study of the impact of a prior commitment to the ratee. *Journal of Applied Psychology, 67*(6), 873-876.

Betz, D. (1974). Rhythmic oscillations influencing grades in oral exams. *Psychologie in Erziehung und Unterricht, 21*(1), 1-14.

Biernat, M., Manis, M., & Kobrynowicz, D. (1997). Simultaneous assimilation and contrast effects in judgments of self and others. *Journal of Personality and Social Psychology, 73*(2), 254-269.

Birkel, P. (1978). *Mündliche Prüfungen: Zur Objektivität und Validität der Leistungsbeurteilung*. Bochum: Verlag Ferdinand Kamp.

Birkel, P. (1984a). Beurteilung mündlicher Prüfungsleistungen. In K. A. Heller (Ed.), *Leistungsdiagnostik in der Schule* (pp. 229-236). Bern: Verlag Hans Huber.

Birkel, P. (1984b). Schriftliche Prüfungen durch Klassenarbeiten. In K. A. Heller (Ed.), *Leistungsdiagnostik in der Schule* (pp. 237-244). Bern: Verlag Hans Huber.

Bless, H., Schwarz, N., & Wänke, M. (2003). The size of context effects in social judgment. In J. P. Forgas & K. D. Williams (Eds.), *Social judgments: Implicit and explicit processes.* (pp. 180-197): Cambridge University Press.

Bodenhausen, G. V., & Macrae, C. N. (1998). Stereotype activation and inhibition. In R. S. J. Wyer (Ed.), *Stereotype activation and inhibition.* (pp. 1-52): Lawrence Erlbaum Associates, Publishers.

Buda, R. (1986). *The effect of expectations, perceived purpose and rating format on performance ratings.*, Univ Microfilms International.

Buunk, B. P., Collins, R. L., Taylor, S. E., & Van Yperen, N. W. (1990). The affective consequences of social comparison: Either direction has its ups and downs. *Journal of Personality & Social Psychology, 59*(6), 1238-1249.

Buunk, B. P., & Mussweiler, T. (2001). New directions in social comparison research. *European Journal of Social Psychology, 31*(5), 467-475.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior & Human Decision, 79*(2), 115-153.

Damisch, L. (2004). *Daumen drücken für den Vorgänger? Der Einfluss von Vergleichen auf das Kampfrichterurteil im Kunstturnen.* Bayerische Julius-Maximilians-Universität Würzburg, Würzburg.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., et al. (1998). Imaging unconscious semantic priming. *Nature, 395*(6702), 597-600.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality & Social Psychology, 56*(1), 5-18.

Devine, P. G., Hirt, E. R., & Gehrke, E. M. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality and Social Psychology, 58*(6), 952-963.

Dijksterhuis, A., Spears, R., Postmes, T., Stapel, D., Koomen, W., Knippenberg, A. v., et al. (1998). Seeing one thing and doing another: Contrast effects in automatic behavior. *Journal of Personality & Social Psychology, 75*(4), 862-871.

Dipboye, R. L., Fromkin, H. L., & Wiback, K. (1975). Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. *Journal of Applied Psychology, 60*(1), 39-43.

Downs, A. C., & Lyons, P. M. (1991). Natural observations of the links between attractiveness and initial legal judgments. *Personality and Social Psychology Bulletin, 17*(5), 541-547.

Dunning, D., & Hayes, A. F. (1996). Evidence for egocentric comparison in social judgment. *Journal of Personality & Social Psychology, 71*(2), 213-229.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but . . .: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*(1), 109-128.

Edwards, K. (1987). Effects of sex and glasses on attitudes toward intelligence and attractiveness. *Psychological Reports, 60*, 590.

Eiser, J. R. (1990). *Social judgment.*: Brooks/Cole Publishing Co.

Englich, B., & Mussweiler, T. (2001). Sentencing under uncertainty: Anchoring effects in the courtroom. *Journal of Applied Social Psychology, 31*(7), 1535-1551.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences.* London: Lawrence Erlbaum.

Evans, J. S. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology, Vol. 64*(3), 391-397.

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology.* (pp. 97-116). New York, NY, US: Guilford.

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73*(3), 421-435.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117-140.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*(4), 659-676.

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom--A cognitive-environmental approach. *Organizational Behavior & Human Decision Processes, Vol 88(1)*, 527-561.

Fiedler, K., Walther, E., & Nickel, S. (1999). The auto-verification of social hypotheses: Stereotyping and the power of sample size. *Journal of Personality & Social Psychology, Vol 77(1)*, 5-18.

Fischer, G. H. (1996). IRT-Modelle als Forschungsinstrumente der Differntiellen Psychologie. In K. Pawlik (Ed.), *Grundlagen und Methoden der Differntiellen Psychologie* (pp. 673-729). Göttingen: Hogrefe.

Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model: Ten years later. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology.* (pp. 231-254): Guilford Press.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition (2nd ed.).* Mcgraw-Hill Book Company.

Ford, T. E., & Thompson, E. P. (2000). Preconscious and postconscious processes underlying construct accessibility effects: An extended search model. *Personality & Social Psychology Review, 4*(4), 317-336.

Forsythe, S., Drake, M. F., & Cox, C. E. (1985). Influence of applicant's dress on interviewer's selection decisions. *Journal of Applied Psychology, 70*(2), 374-378.

Funder, D. C. (2003). Toward a social psychology of person judgments: Implications for person perception accuracy and self-knowledge. In J. P. Forgas & K. D. Williams (Eds.), *Social judgments: Implicit and explicit processes.* (pp. 115-133): Cambridge University Press.

Gadenne, V., & Oswald, M. (1986). Entstehung und Veraenderung von Bestaetigungstendenzen beim Testen von Hypothesen. *Zeitschrift fuer Experimentelle und Angewandte Psychologie, 33*(3), 360-374.

Gati, I., & Tversky, A. (1987). Recall of common and distinctive features of verbal and pictorial stimuli. *Memory & Cognition, 15*(2), 97-100.

Gilbert, D. T., Giesler, R. B., & Morris, K. A. (1995). When comparisons arise. *Journal of Personality & Social Psychology, 69*(2), 227-236.

Grice, H. P. (2002). Logic and conversation. In D. J. Levitin (Ed.), *Foundations of cognitive psychology: Core readings.* (pp. 719-732): MIT Press.

Gulliksen, H. (1950). *Theory of mental tests.*: Wiley.

Hakmiller, K. (1966). Threat as a determinant of social comparison. *Journal of Experimental Social Psychology, Suppl. 1*, 32-39.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93*(3), 258-268.

Helson, H. (1964). *Adaptation-level theory: an experimental and systematic approach to behavior.*: Harper and Row: New York.

Herr, P. M. (1986). Consequences of priming: Judgment and behavior. *Journal of Personality & Social Psychology, 51*(6), 1106-1115.

Herr, P. M., Sherman, S. J., & Fazio, R. H. (1983). On the consequences of priming: Assimilation and contrast effects. *Journal of Experimental Social Psychology, 19*(4), 323-340.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles.* (pp. 133-168): Guilford Press.

Higgins, E. T., & King, G. (1981). Accessibility of social constructs: Information-processing consequences of individual and contextual variability. In N. Cantor & J. F. Kihlstrom (Eds.), *Personality, cognition and social interaction* (pp. 69-121). Hillsdale, NJ: Erlbaum.

Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology, 13*(2), 141-154.

Highhouse, S., & Gallo, A. (1997). Order effects in personnel decision making. *Human Performance, 10*(1), 31-46.

Houston, D. A., Sherman, S. J., & Baker, S. M. (1989). The influence of unique features and direction of comparison on preferences. *Journal of Experimental Social Psychology, 25*(2), 121-141.

Houston, D. A., Sherman, S. J., & Baker, S. M. (1991). Feature matching, unique features, and the dynamics of the choice process: Predecision conflict and postdecision satisfaction. *Journal of Experimental Social Psychology, 27*(5), 411-430.

Ilgen, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to perfromance appraisal processes. *Academy of Management Review, 10*, 311-321.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136-153.

Kahneman, D., & Miller, D. T. (2002). Norm theory: Comparing reality to its alternatives. In T. Gilovich & D. Griffin (Eds.), *Heuristics and biases: The psychology of intuitive judgment.* (pp. 348-366): Cambridge University Press.

Kernis, M. H., & Wheeler, L. (1981). Beautiful friends and ugly strangers: Radiation and contrast effects in perceptions of same-sex pairs. *Personality and Social Psychology Bulletin, 7*(4), 617-620.

Klayman, J., & Ha, Y.-w. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211-228.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110*(3), 499-519.

Kühnen, U., & Haberstroh, S. (2004). Self-construal activation and focus of comparison as determinants of assimilation and contrast in social comparisons. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 22*(2), 289-310.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin, 129*(4), 522-544.

Lepore, L., & Brown, R. (1997). Category and stereotype activation: Is prejudice inevitable? *Journal of Personality & Social Psychology, 72*(2), 275-287.

Manis, M., Nelson, T. E., & Shedler, J. (1988). Stereotypes and social judgment: Extremity, assimilation, and contrast. *Journal of Personality & Social Psychology, 55*(1), 28-36.

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory and Cognition, 24*(2), 235-249.

Martin, L. L. (1986). Set/reset: Use and disuse of concepts in impression formation. *Journal of Personality & Social Psychology, 51*(3), 493-504.

Martin, L. L., & Achee, J. W. (1992). Beyond accessibility: The role of processing objectives in judgment. In L. L. Martin & A. Tesser (Eds.), *Construction of social judgments.* (pp. 195-216): Lawrence Erlbaum Associates, Inc.

Miller, C. T. (1984). Self-schemas, gender, and social comparison: A clarification of the related attributes hypothesis. *Journal of Personality & Social Psychology, 46*(6), 1222-1229.

Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology, 70*(1), 72-84.

Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective*. Boston: Allyn & Bacon.

Murphy, K. R., Gannett, B. A., Herr, B. M., & Chen, J. A. (1986a). Effects of subsequent performance on evaluations of previous performance. *Journal of Applied Psychology, 71*(3), 427-431.

Murphy, K. R., Gannett, B. A., Herr, B. M., & Chen, J. A. (1986b). Effects of subsequent performance on evaluations of previous performance. *Journal of Applied Psychology,, 71*(3), 427-431.

Murphy, N. A., Hall, J. A., & LeBeau, L. S. (2001). Who's smart? Beliefs about the expression of intelligence in social behavior. *Representative Research in Social Psychology, 25*, 34-42.

Mussweiler, T. (2001a). Focus of comparison as a determinant of assimilation versus contrast in social comparison. *Personality and Social Psychology Bulletin, 27*(1), 38-47.

Mussweiler, T. (2001b). 'Seek and ye shall find': Antecedents of assimilation and contrast in social comparison. *European Journal of Social Psychology, 31*(5), 499-509.

Mussweiler, T. (2003a). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review, 110*(3), 472-489.

Mussweiler, T. (2003b). 'Everything is relative': Comparison processes in social judgment: The 2002 Jaspars Lecture. *European Journal of Social Psychology, 33*(6), 719-733.

Mussweiler, T., & Bodenhausen, G. V. (2002). I know you are, but what am I? Self-evaluative consequences of judging in-group and out-group members. *Journal of Personality & Social Psychology, 82*(1), 19-32.

Mussweiler, T., & Rüter, K. (2003). What Friends Are For! The Use of Routine Standards in Social Comparison. *Journal of Personality & Social Psychology, 85*(3), 467-481.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35*(2), 136-164.

Mussweiler, T., & Strack, F. (2000a). Consequences of social comparison: Selective accessibility, assimilation, and contrast. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 253-270). New York, NY, US: Kluwer Academic / Plenum Publishers.

Mussweiler, T., & Strack, F. (2000b). The "relative self": Informational and judgmental consequences of comparative self-evaluation. *Journal of Personality & Social Psychology, 79*(1), 23-38.

Mussweiler, T., & Strack, F. (2000c). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality & Social Psychology, 78*(6), 1038-1052.

Ostrom, T. M., & Upshaw, H. S. (1968). Psychological perspective and attitude change. In A. C. Greenwald, T. C. Brock & T. M. Ostrom (Eds.), *Psychological foundations of attitudes* (pp. 217-242). New York, NY, USA: Academic Press.

Oswald, M. E. (1993). Hypothesentesten: Suche und Verarbeitung hypothesenkonformer und hypothesenkonträrer Informationen. In W. Hell, K. Fielder & G. Gigerenzer (Eds.), *Kognitive Täuschungen: Fehl-Leistungen und Mechanismen des Urteilens, Denkens und Erinnerns* (pp. 189-211). Heidelberg: Spektrum Akademischer Verlag.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*(6), 407-418.

Petty, R. E., & Wegener, D. T. (1993). Flexible correction processes in social judgment: Correcting for context-induced contrast. *Journal of Experimental Social Psychology, 29*(2), 137-165.

Petzhold, P., & Haubensak, G. (2001). Higher order sequential effects in psychophysical judgments. *Perception & Psychophysics, Vol 63(6)*, 969-978.

Plessner, H. (1997). *Urteilsverzerrungen bei Kampfrichtern im Kunstturnen*. Aachen: Shaker Verlag.

Plessner, H. (1999). Expectation biases in gymnastics judging. *Journal of Sport & Exercise Psychology, 21*(2), 131-144.

Plessner, H., & Betsch, T. (2001). Sequential effects in important referee decisions: The case of penalties in soccer. *Journal of Sport & Exercise Psychology, 23*(3), 254-259.

Raymond, M. R., Webb, L. C., & Houston, W. M. (1991). Correcting performance-rating errors in oral examinations. *Evaluation and the Health Professions, 14*(1), 100-122.

Reis, H. T., Wilson, I. M., Monestere, C., Bernstein, S., Clark, K., Seidl, E., et al. (1990). What is smiling is beautiful and good. *European Journal of Social Psychology, 20*(3), 259-267.

Rosenthal, R. (1991). Teacher expectancy effects: A brief update 25 years after the pygmalion experiment. *Journal of Research in Education, 1*, 3-12.

Ruys, K. I. (2004). *The influence of self and similarity on immediate affective impressions of people.* Unpublished dissertation, Universiteit van Amsterdam, Amsterdam.

Sanbonmatsu, D. M., Posavac, S. S., Kardes, F. R., & Mantel, S. P. (1998). Selective hypothesis testing. *Psychonomic Bulletin & Review, 5*(2), 197-220.

Schnotz, W. (1971). Estimates of task difficulties by teachers. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, Vol. 3*(2), 106-120.

Schuh, A. J. (1978). Contrast effect in the interview. *Bulletin of the Psychonomic Society, 11*(3), 195-196.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. In M. P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 26.* (pp. 123-162): Academic Press.

Schwarz, N., & Bless, H. (1992a). Constructing reality and its alternatives: An inclusion/exclusion model of assimilation and contrast effects in social judgment. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 217-245). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Schwarz, N., & Bless, H. (1992b). Scandals and the public's trust in politicians: Assimilation and contrast effects. *Personality & Social Psychology Bulletin, 18*(5), 574-579.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality & Social Psychology, Vol 45(3)*, 513-523.

Schwarz, N., & Strack, F. (1999). Reports of subjective well-being: Judgmental processes and their methodological implications. In D. D. Kahneman, Ed (Ed.), *Well-being: The foundations of hedonic psychology.* (pp. 61-84). New York, NY, US: Russell Sage Foundation.

Sherif, M. (1936). *The psychology of social norms.* New York, NY, USA: Harper.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22*(2), 93-121.

Smith, E. R. (1994). Procedural knowledge and processing strategies in social cognition. In R. S. J. Wyer & T. K. Srull (Eds.), *Handbook of social cognition, Vol. 1: Basic processes; Vol. 2: Applications (2nd ed.).* (pp. 99-151): Lawrence Erlbaum Associates, Inc.

Smith, E. R., & Mackie, D. M. (1995). *Social Psychology*. New York, NY, USA: Worth Publishers.

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99*(1), 3-21.

Smither, J. W., Reilly, R. R., & Buda, R. (1988). Effect of prior performance information on ratings of present performance: Contrast versus assimilation revisited. *Journal of Applied Psychology, 73*(3), 487-496.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36*(11), 1202-1212.

Snyder, M., & Uranowitz, S. W. (1978). Reconstructing the past: Some cognitive consequences of person perception. *Journal of Personality & Social Psychology, 36*(9), 941-950.

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality & Social Psychology, 37*(10), 1660-1672.

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review, 96*(1), 58-83.

Stapel, D. A., & Koomen, W. (2001a). The impact of interpretation versus comparison mindsets on knowledge accessibility effects. *Journal of Experimental Social Psychology, 37*(2), 134-149.

Stapel, D. A., & Koomen, W. (2001b). Let's not forget the past when we go to the future: On our knowledge of knowledge accessibility. In G. B. Moskowitz (Ed.), *Cognitive social psychology: The Princeton Symposium on the Legacy and Future of Social Cognition.* (pp. 229-246): Lawrence Erlbaum Associates, Publishers.

Stapel, D. A., Koomen, W., & van der Pligt, J. (1996). The referents of trait inferences: The impact of trait concepts versus actor-trait links on subsequent judgments. *Journal of Personality & Social Psychology, 70*(3), 437-450.

Stapel, D. A., Koomen, W., & Zeelenberg, M. (1998). The impact of accuracy motivation on interpretation, comparison, and correction processes: Accuracy X knowledge accessibility effects. *Journal of Personality & Social Psychology, Vol 74(4)*, 878-893.

Stapel, D. A., & Schwinghammer, S. A. (2004). Defensive social comparisons and the constraints of reality. *Social Cognition, 22*(1), 147-167.

Strack, F. (1992). The different routes to social judgments: Experiential versus informational strategies. In L. L. Martin & A. Tesser (Eds.), *Construction of social judgments.* (pp. 249-275): Lawrence Erlbaum Associates, Inc.

Strack, F., & Martin, L. (1987). Thinking, judging, and communicating: A process account for context effects in attitude surveys. In H. J. Hippler, N. Schwarz & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 123-148). New York, NY, USA: Springer-Verlag.

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology, 18*(5), 429-442.

Strack, F., Schwarz, N., & Gschneidinger, E. (1985). Happiness and reminiscing: The role of time perspective, affect, and mode of thinking. *Journal of Personality & Social Psychology, 49*(6), 1460-1469.

Suls, J., Martin, R., & Wheeler, L. (2002). Social comparison: Why, with whom, and with what effect? *Current Directions in Psychological Science, 11*(5), 159-163.

Suls, J., & Wheeler, L. (2000). A selective history of classic and neo-social comparioson theory. In J. Suls & L. Wheeler (Eds.), *Handbook of Social Comparison: Theory and Research* (pp. 3-19). New York: Kluwer Academic / Plenum Publishers.

Sumer, H. C., & Knight, P. A. (1996). Assimilation and contrast effects in performance ratings: Effects of rating the previous performance on rating subsequent performance. *Journal of Applied Psychology, 81*(4), 436-442.

Taylor, S. E., & Lobel, M. (1989). Social comparison activity under threat: Downward evaluation and upward contacts. *Psychological Review, 96*(4), 569-575.

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology, Vol. 21: Social psychological studies of the self: Perspectives and programs.* (pp. 181-227): Academic Press, Inc.

Thornton, D., & Arrowood, A. J. (1966). Self-evaluation, self-enancement, and the locus of social comparison. *Journal of Experimental Social Psychology, Suppl. 1*, 40-48.

Traut-Mattausch, E., Schulz-Hardt, S., Greitemeyer, T., & Frey, D. (2004). Expectancy confirmation in spite of disconfirming evidence: The case of price

increases due to the introduction of the Euro. *European Journal of Social Psychology, 34*(6), 739-760.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93*(3), 239-257.

Trope, Y., & Liberman, A. (1996). Social hypothesis testing: Cognitive and motivational mechanisms. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles:* (pp. 239-270). New York, NY, US: The Guilford Press.

Trope, Y., & Thompson, E. P. (1997). Looking for truth in all the wrong places? Asymmetric search of individuating information about stereotyped group members. *Journal of Personality and Social Psychology, 73*(2), 229-241.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352.

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization.* (pp. 79-98). New York: Lawrence Erlbaum.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89*(2), 123-154.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.

Wedell, D. H., Parducci, A., & Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *American Journal of Psychology, 102*(2), 233-248.

Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. *Journal of Personality & Social Psychology, 68*(1), 36-51.

Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 141-208): Academic Press.

Wegener, D. T., Petty, R. E., & Dunn, M. (1998). The metacognition of bias correction: Naive theories of bias and the flexible correction model. In G. Lories (Ed.), *Metacognition: Cognitive and social dimensions.* (pp. 202-227): Sage Publications, Inc.

Wexley, K. N., Yukl, G. A., Kovacs, S. Z., & Sanders, R. E. (1972). Importance of contrast effects in employment interviews. *Journal of Applied Psychology, Vol. 56*(1), 45-48.

Wheeler, L., Koestner, R., & Driver, R. E. (1982). Related attributes in the choice of comparison others: It's there, but it isn't all there is. *Journal of Experimental Social Psychology, 18*(6), 489-500.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35*(3), 521-551.

Williams, K. J., Cafferty, T. P., & DeNisi, A. S. (1990). Effects of appraisal salience on recall and ratings. *Organizational Behavior and Human Decision, 46*, 217-239.

Wills, T. A. (1981). Downward comparison principles in social psychology. *Psychological Bulletin, 90*(2), 245-271.

Wood, J. V., Taylor, S. E., & Lichtman, R. R. (1985). Social comparison in adjustment to breast cancer. *Journal of Personality & Social Psychology, 49*(5), 1169-1183.

Zuckerman, M., Knee, C. R., Hodgins, H. S., & Miyake, K. (1995). Hypothesis confirmation: The joint effect of positive test strategy and acquiescence response set. *Journal of Personality and Social Psychology, 68*(1), 52-60.

Appendices

Appendices chapter 2

Appendix 2.A: Priming task (instructions, comparison task)

Liebe/r Versuchsteilnehmer/in,

bei der folgenden Aufgabe sollen Sie zwei Bilder miteinander vergleichen. Diese sehen Sie auf der nächsten Seite. Ihre Aufgabe besteht darin, so viele Ähnlichkeiten wie möglich zwischen diesen beiden Bildern zu finden. Diese sollen Sie auf der nächsten Seite aufschreiben.

VERSION A:

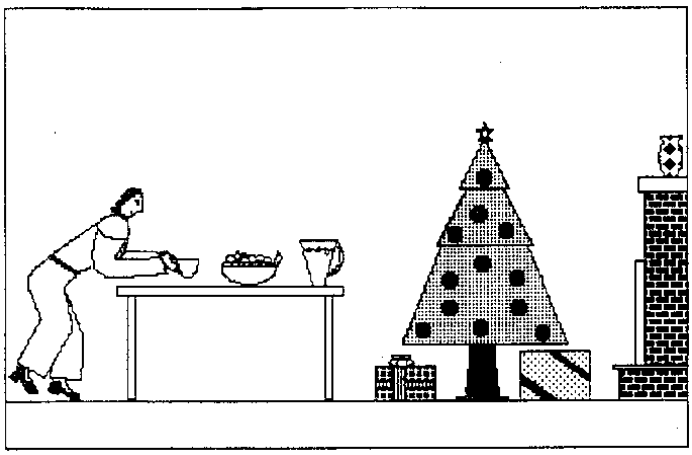
Versuchen Sie dabei bitte, alle Ähnlichkeiten aufzulisten, die es zwischen den beiden Bildern gibt.
Nehmen Sie sich dazu ruhig ein paar Minuten Zeit.

VERSION B:

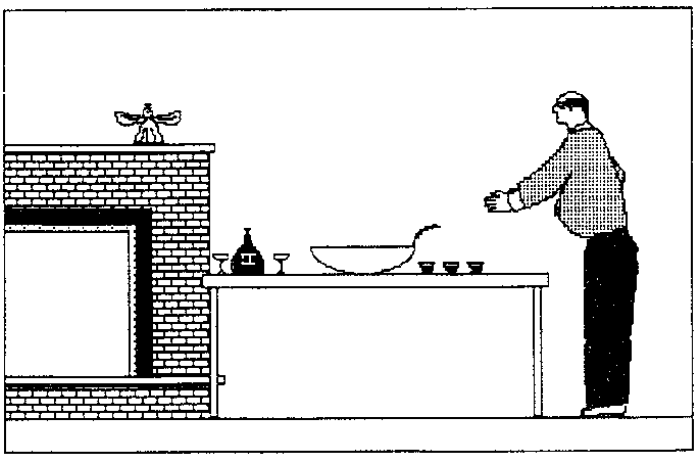
Versuchen Sie dabei bitte, alle Unterschiede aufzulisten, die es zwischen den beiden Bildern gibt.
Nehmen Sie sich dazu ruhig ein paar Minuten Zeit.

Wenn Sie noch Fragen haben, wenden Sie sich bitte an die Versuchsleitung!

Ansonsten blättern Sie bitte um und bearbeiten Sie die Aufgabe.



A



B

VERSION A: **Worin sind sich diese beiden Bilder ähnlich?**
VERSION B: **Worin unterscheiden sich diese beiden Bilder?**

Melden Sie sich bitte bei der Versuchsleitung, wenn Sie mit der Bearbeitung der Aufgabe fertig sind

Appendix 2.B: Written protocols (instructions, high standard, low standard, target)

Liebe/r Versuchsteilnehmer/in,

in dieser Studie geht es um das Testen von Allgemeinwissen. Sie sollen dabei in die Rolle eines Prüfers schlüpfen und Prüfungsleistungen bewerten.

Stellen Sie sich folgende Situation vor:

Eine Schule in Baden-Württemberg hat aufgrund der PISA-Schlagzeilen beschlossen für die Schüler des Abschlussjahrgangs in jedem Jahr einen Preis für das beste Allgemeinwissen auszusetzen. Dadurch soll die Motivation zu lernen erhöht werden.

Ein Preis soll für das Wissen in den Bereichen Chemie, Geographie, Mathematik und Physik vergeben werden. Für diesen Preis sind zwei Schüler vorgeschlagen worden. Beide haben jeweils 8 Fragen aus den genannten Bereichen in einem schriftlichen Test vorgelegt bekommen. Sie sollen nun als objektiver Prüfer die Antworten der beiden Schüler lesen, die Prüfungsleistung benoten und am Ende entscheiden, welcher der beiden Schüler den Preis aufgrund seines Allgemeinwissens erhalten soll.

Bitte lesen Sie die Antworten der beiden Schüler aufmerksam durch und vergeben Sie für jeden eine Note von 1,0 – 5,0. Bedenken Sie, dass Sie nicht nur die Leistung der beiden Schüler beurteilen sollen. Zusätzlich sollen Sie am Ende entscheiden, welcher der beiden Schüler den Preis aufgrund seines Allgemeinwissens erhalten sollte.

Bitte blättern Sie bei der Bearbeitung dieses Fragebogens nicht zurück!

Falls Sie Fragen haben, wenden Sie sich bitte jederzeit an die Versuchsleitung.

Blättern Sie jetzt bitte um, und beginnen Sie mit der Bewertung des ersten Schülers.

Die erste Schülerin heißt Christina und ist 18 Jahre alt.

1) Wie lautet die Abkürzung für die Organisation erdölexportierender Länder?

OPEC ist die Abkürzung für die „Organisation der Erdöl exportierenden Länder“ und steht eigentlich für den englischen Term „Organization of the Petroleum Exporting Countries“. Diese wurde im Jahre 1960 gegründet und ihr gehören z.B. folgende Staaten an: Kuwait, Saudi-Arabien und Vereinigte Arabische Emirate. Die OPEC war zunächst eine Schutzorganisation gegen die Ölkonzerne, mit denen Fördermengen und feste Rohölpreise ausgehandelt wurden. Zur autonomen Preisfestsetzung (und zum Einsatz des Erdöls als politisches Druckmittel) ging die OPEC über, als die Ölkonzerne bei Verhandlungen 1973 zögerten, die Forderungen nach höherem Inflationsausgleich und Anhebung der Listenpreise zu akzeptieren. Es kam zum ersten „Ölpreisschock“, der bis Anfang 1974 zu einer Vervierfachung des Rohölpreises führte. Seither ist der OPEC-Anteil an der Weltförderung von 47,8% (= 1524 Millionen Tonnen) im Jahre 1979 auf 38,5% im Jahre 1991 zurückgegangen.

2) Für welche Erfindung wurde Bunsen berühmt?

Robert Wilhelm Bunsen, geboren 1811 in Göttingen, hat in seiner wissenschaftlichen Karriere einige wichtige Erfindungen gemacht. Er fand zum Beispiel ein Gegenmittel bei Arsenvergiftungen und eine neue Methode um Starkstrom zu gewinnen. Doch am meisten bekannt geworden ist er für den Bunsenbrenner, den jedes Kind aus dem Chemieunterricht kennt. Der Bunsenbrenner wurde 1855 erfunden und ist ein Leuchtgasbrenner, der aus Düse, Luftregler und Brennrohr besteht. Das aus der Düse in das Brennrohr einströmende Gas saugt über eine verstellbare Öffnung Luft an. Je größer die Öffnung, desto heißer die Flamme, je kleiner, desto heller und kühler wird die Flamme. Zusammen mit Gustav Robert Kirchhoff und einem Bunsenbrenner als Hilfsmittel entwickelte Robert Wilhelm Bunsen im Jahre 1859 die Spektralanalyse. Alle atomaren Gase emittieren nach der energetischen Anregung bestimmte Spektrallinien in einer charakteristischen Kombination, die eine analytische Erfassung auf diesem Wege ermöglicht. Die Spektralanalyse kann seither auch auf Sonnenlicht und selbstleuchtende Himmelskörper angewendet werden.

3) Was ist das Fachwort für Erdanziehungskraft?

Das Fachwort für Erdanziehungskraft – also die Anziehungskraft zwischen der Erde und den in ihrer Nähe befindlichen Körpern – ist Gravitation. Mit Gravitation ist aber auch die Schwerkraft bzw. Anziehungskraft zwischen Körpern im allgemeinen gemeint. Gravitation ist die universelle Eigenschaft aller Körper sich gegenseitig anzuziehen. Dies wurde schon von Isaac Newton in seinem Gravitationsgesetz formuliert, dass die Kraft angibt, mit der sich zwei Massen anziehen: Sie beträgt die Masse des ersten Körpers mal der Masse des zweiten Körpers geteilt durch das Quadrat des Abstands beider Körper. Diesen Wert muss man noch mit der sogenannten Gravitationskonstante multiplizieren. Die wesentlichen Eigenschaften lassen sich aber auch so erkennen: Die Anziehungskraft zweier Körper ist abhängig von ihrer Masse und ihrer Entfernung voneinander. Warum jedoch Körper eine Gravitationswirkung haben, ist fast schon eine philosophische Frage. Albert Einsteins Vorstellung davon kann man sich - sehr vereinfacht - so vorstellen: Jeder Körper verursacht eine "Mulde" in der Raumzeit. Je massereicher ein Körper ist, desto größer ist diese Mulde und desto stärker werden andere Körper angezogen: Sie fallen einfach in diese Mulde hinein.

4) Wie heißen die Achsen im zweidimensionalen Koordinatensystem?

Der Begriff des Koordinatensystems kommt aus der Seefahrt und zwar im Zusammenhang mit der Bestimmung der raum-zeitlichen Position eines Schiffes. Mittels eines Sextanten, einem Instrument zum Freihandmessens von Gestirnshöhen, werden über die Bestimmung der Winkel die Koordinaten des Schiffes gemessen und in ein Koordinatensystem eingetragen. Die Achsen in einem zweidimensionalen Koordinatensystem heißen α-Achse und β-Achse. Allerdings wird normalerweise auch die dritte Dimension über die sogenannten γ-Achse erfasst. Bei der Arbeit mit dem Sextanten sollte man allerdings den Kippfehler des Horizontspiegels beachten. Dieser Fehler tritt auf, wenn der Horizontspiegel nicht genau senkrecht steht. Außerdem kann es zu einem Indexfehler kommen, wenn die Einheiten zwischen den Achsen nicht gleichförmig gewählt sind.

Blättern Sie bitte um !

5) Zu welcher Inselgruppe gehört Gran Canaria?

Gran Canaria gehört zur Inselgruppe der Canaren. Die Inselgruppe liegt vor der Nordwestküste Afrikas. Zu ihr zählen 7 größere Inseln, nämlich Gran Canaria, Lanzarote, Fuerteventura, Gomera, Hierro, Teneriffa und La Palma sowie einige kleinere Inseln, wie zum Beispiel Alegranza, Graciosa oder Montaña Clara. Die Landschaftstypen variieren von Insel zu Insel. Vorherrschend sind Felsklippen, Vulkangestein, Bergketten und Waldgebiete. Auf den nördlichen Inseln ist das Klima subtropisch, auf den südlichen trockener mit noch höheren Temperaturen. Niederschläge sind sehr selten. Santa Cruz de Tenerife ist die Hauptstadt für Teneriffa, La Palma, Gomena und Hierro und hat ca. 200.000. Einwohner. Las Palmas de Gran Canaria ist die Hauptstadt für Gran Canaria, Fuerteventura und Lanzarote und hat ca. 355.000 Einwohner. Arabische Seefahrer erreichten die Inselgruppe im 12. Jahrhundert. Der französische Seefahrer Jean de Behténcourt begann 1402 mit der Eroberung der Inseln und wurde 1404 von Heinrich III. zum König der Kanaren gekrönt. Die Inseln wurden durch einen Vertrag im Jahre 1479 spanische Besitzungen. 1496 wurden die Inseln der der spanischen Krone einverleibt, wenig später machten die Schiffe von Christoph Kolumbus hier Station auf ihrer Entdeckungsreise in die Neue Welt. Seit 1975 ist König Juan Carlos I. Staatsoberhaupt.

6) Wie berechnet man einen Mittelwert?

Der Mittelwert oder das arithmetische Mittel ist eines der Maße der zentralen Tendenz. Die Maße der zentralen Tendenz einer Verteilung haben als statistische Kennwerte die Funktion, über spezielle Eigenschaften einer Verteilung summarisch Auskunft zu geben. Dabei versuchen die Maße der zentralen Tendenz einen Wert anzugeben, der die gesamte Verteilung am besten repräsentiert. Die gebräuchlichsten Maße der zentralen Tendenz sind der Modalwert, der Medianwert und das arithmetische Mittel bzw. Mittelwert. Der Mittelwert als der gebräuchlichste Kennwert wird berechnet, indem die Summe aller Werte durch die Anzahl aller Werte dividiert wird. Der Mittelwert hat die Eigenschaft, dass die Summe der quadratischen Abweichungen aller Werte vom Mittelwert ein Minimum ergibt. Die Berechnung des Mittelwerts setzt voraus, dass das untersuchte Merkmal mindestens intervallskaliert ist. Im Gegensatz zum Mittelwert ist der Modalwert der Wert einer Verteilung, der am häufigsten vorkommt und der Median der Wert, von dem alle übrigen Werte im Durchschnitt am wenigsten Abweichen.

7) Was ist das Periodensystem der Elemente?

In den Jahren 1868/69 ordneten der russische Chemiker Dimitri Iwanowitsch Mendelejew und der deutsche Lothar Meyer unabhängig voneinander die damals bekannten Elemente in einem System an, in dem die Reihe der Elemente in der Weise in Perioden zerlegt ist, dass Elemente mit ähnlichen Eigenschaften in Gruppen zusammengefasst sind. Dieses System ist als Periodensystem der Elemente (kurz PSE) bekannt. Die nach ihren Kernladungszahlen (= Ordnungszahlen) geordneten Elemente zeigen eine sich periodisch wiederholende Ähnlichkeit von Eigenschaften. Mendelejew war aufgrund des von ihm aufgestellten Periodensystem in der Lage, die Existenz der damals noch nicht bekannten Elemente Gallium ('Eka-Aluminium') und Germanium ('Eka-Silicium') vorherzusagen.Die waagrechten Zeilen des Periodensystems heißen Perioden. Die senkrechten Spalten des Periodensystems heißen Gruppen. Es wird zwischen Haupt- und Nebengruppen unterschieden.

8) Was bedeutet die Abkürzung UKW?

UKW ist die Abkürzung für Ultrakurzwelle. Damit wird ein weltweiter Frequenzbereich für die Übertragung von Stereo-Radioprogrammen in analoger Technik beschrieben. Dieser Frequenzbereich liegt zwischen 87,5 und 108 Mhz. Schon zu Zeiten der Kurzwelle wurde entdeckt, dass es noch darüber hinaus höhere Frequenzen gibt. Diese galten lange Zeit als nicht nutzbar. Erst mit der Weiterentwicklung und Verfeinerung der Röhrentechnik gelang es, den Sendern eine verwertbare Ausgangsleistung und den Empfängern eine passable Empfindlichkeit zu geben. Dies geschah im Zeitraum kurz vor und im Zweiten Weltkrieg. Nach Ende des Zweiten Weltkrieges und der Zulassung des Amateurfunks in Deutschland durch die Alliierten begannen Funkamateure sofort damit, mit Röhren aus Beständen der Streitkräfte Versuche im UKW-Bereich zu machen. Mit Einführung der Transistoren wurden die Geräte kleiner, die Empfindlichkeit und die Sendeleistung bekamen nochmals einen Leistungsschub, und bald konnte man schon Autoradios mit UKW-Bereich kaufen.

Blättern Sie bitte um !

Die erste Schülerin heißt Christina und ist 18 Jahre alt.

1) Wie lautet die Abkürzung für die Organisation erdölexportierender Länder?

Die Abkürzung lautet UNO. Die Organisation der Vereinten Nationen (UNO) ist die einzige universelle Organisation. Ihre zentralen Themen sind Sicherheit und Frieden, Menschenrechte, Förderung der Wohlfahrt, Abbau sozialer Gegensätze und humanitäre Hilfe sowie Schutz der natürlichen Lebensgrundlagen. In ihrer Charta steht, dass die Völker der vereinten Nationen Grundsätze annehmen und Verfahren einführen wollen, die gewährleisten, dass Waffengewalt nur noch im gemeinsamen Interesse angewendet wird, und außerdem internationale Einrichtungen in Anspruch zu nehmen, um den wirtschaftlichen und sozialen Fortschritt aller Völker zu fördern. Frieden und wirtschaftlicher Wohlstand sind in unserer Welt in höchstem Maße vom Erdöl abhängig. Deshalb versucht die UNO auch, ihre Ziele - wie wirtschaftlichen Wohlstand und Frieden - durch die Kontrolle der Abgabemengen und –preise von Erdöl weltweit zu erreichen.

2) Für welche Erfindung wurde Bunsen berühmt?

Robert Wilhelm Bunsen, geboren 1811 in Göttingen, hat in seiner wissenschaftlichen Karriere einige wichtige Erfindungen gemacht. Er fand zum Beispiel ein Gegenmittel für Schlangengift und eine neue Methode um Strom zu speichern. Doch am meisten bekannt geworden ist er für die Erfindung des Mikroskops. Zusammen mit dem holländischen Brillenmacher Zacharias Jansen erfand er das zusammengesetzte Mikroskop sowie später auch das Teleskop um das Jahr 1850. Die ersten Mikroskope bestanden aus zwei Sammellinsen mit denen man auch Sonnenstrahlen bündeln konnte – deshalb erhielt das Mikroskop auch den Spitznamen Bunsen-Brenner. Erst zu Beginn des 19. Jahrhunderts mauserte sich das Mikroskop vom Apparat der Volksbelustigung zum wichtigsten wissenschaftlichen Instrument in der Medizin und in den Naturwissenschaften. Ein Mikroskop ist ein Instrument, das es erlaubt, sehr kleine Objekte vergrößert anzusehen. Der Name kommt aus dem griechischen von mikrós = klein und skopein = betrachten. Es wird zwischen folgenden Mikroskoptypen unterschieden: Dem Lichtmikroskop, den Elektronenmikroskopen und den Rastersondenmikroskopen

3) Was ist das Fachwort für Erdanziehungskraft?

Das Fachwort für Erdanziehungskraft ist Zentrifugalkraft. Darunter versteht man die Kraft, die ein Beobachter auf einer Kreisbahn verspürt. Sie ist vom Mittelpunkt nach außen gerichtet, ist vom Betrag her gleich der Zentripetalkraft und selbiger entgegen gerichtet. Durch die Drehung der Erde werden wir also von der Zentrifugalkraft nach außen in das All gedrückt. Die Zentrifugalkraft ist eine Kraft, die der Beschleunigung des Bezugssystems entgegengerichtet ist. Man bezeichnet sie als Trägheitskraft, sie ist im beschleunigten System direkt zu messen. Oft nennt man solche Trägheitskräfte auch Scheinkräfte, da sie im nicht beschleunigten System nicht erkennbar sind. Das beschleunigte Bezugssystem birgt allerdings das Problem, dass die Newtonschen Gesetze nicht mehr uneingeschränkt gelten: Das 3. Gesetz von Newton besagt, dass zu jeder Kraft an einem Körper eine Gegenkraft an einem anderen Körper existieren muss. Dies ist für Scheinkräfte nicht erfüllt, sie haben keine "reactio".

4) Wie heißen die Achsen im zweidimensionalen Koordinatensystem?

Ein Gitter mit zwei Achsen heißt Koordinatensystem. Der Schnittpunkt der Achsen (0/0) heißt Ursprung. Ein Punkt im Koordinatensystem wird als Zahlenpaar (x/y) angegeben. Der x–Wert heißt x-Koordinate, der y–Wert heißt y-Koordinate. Die Koordinaten zeigen dir die genaue Lage zum Ursprung des Systems. Die erste Achse heißt Abszisse, die zweite Ordinate – man spricht auch von x- und y-Achse. Die Bezeichnungen x und y werden zwar häufig verwendet (man spricht dann von einem xy-Koordinatensystem), sind aber nicht obligatorisch. Man könnte an ihrer Stelle irgendwelche anderen Symbole verwenden. Die beiden Achsen, zusammen mit ihren Orientierungen (den Pfeilen) bilden ein (kartesisches oder rechtwinkeliges) Koordinatensystem. Sein Sinn ist es also, uns eine Vorschrift in die Hand zu geben, wie die Position eines Punktes durch zwei Zahlen (den Koordinaten) ausgedrückt werden kann. Dadurch werden viele geometrische Probleme einer rechnerischen Behandlung zugänglich gemacht.

Blättern Sie bitte um !

5) Zu welcher Inselgruppe gehört Gran Canaria?

Gran Canaria befindet sich im Mittelmeer und gehört zu der Inselgruppe der Balearen. Palma de Mallorca, die Hauptstadt der Balearen, ist ca.181 km von Gran Canaria entfernt. Gran Canaria wird eingegrenzt durch die Breitengrade 40°5´17´´ und 39°47´55´´ und durch die Längengrade 3°52´00´´ und 4°24´00´´. Gran Canaria hat eine Fläche von ca. 718,84 km² und eine Küstenlange von ca. 285,70 km. Die Insel ist in Ostwestrichtung etwa 50 km lang und in Nordsüdrichtung durchschnittlich 16 km breit. Gran Canaria hat ca. 69 000 Einwohner und ist damit etwa nur halb so dicht besiedelt wie die Nachbarinseln. Wie fast ganz Spanien wurde einst auch Gran Canaria von den Arabern besetzt. Von 903 an gehörte die Insel zum Kalifat Córdoba und blieb knapp vier Jahrhunderte unter islamischer Herrschaft, bis im Januar 1287 Alfons III. von Aragonien die Insel für das Christentum zurückeroberte. Nach einigen Eroberungen durch die Briten und die Franzosen gehört Gran Canaria sowie die gesamten Balearen heute zu Spanien.

6) Wie berechnet man einen Mittelwert?

Der Mittelwert ist derjenige Wert einer Verteilung, der am häufigsten besetzt ist, bzw. in der grafischen Darstellung einer Verteilung der Wert, bei dem die Verteilung ihr Maximum hat. Handelt es sich um eine diskrete Skala, bei der keine Messwerte zu Kategorien zusammengefasst sind, ist der am häufigsten auftretende Messwert der Mittelwert. Wurden die Messwerte in Kategorien zusammengefasst, gilt die Kategorienmitte der am häufigsten besetzten Kategorie als Mittelwert. Der Mittelwert setzt voraus, dass alle Werte positiv sind, und wird wie folgt berechnet: 1. Auszählen der Messwerte, die aus einer Kategorie benötigt werden, um genau 50% zu erreichen. 2. Division dieser Werte durch die Anzahl der Werte in dieser Kategorie. 3. Multiplikation des Quotienten mit der Kategorienbreite. 4. Addition dieses Wertes zu der unteren Grenze der kritischen Kategorie. Diese Rechenschritte bezeichnet man auch als lineare Interpolation.

7) Was ist das Periodensystem der Elemente?

Das Periodensystem beschreibt den Ablauf des weiblichen Zyklus und geht zurück auf die unabhängigen Untersuchungen des russischen Biologen Dimitri Iwanowitsch Mendelejew und des deutschen Mediziners Lothar Meyer aus den Jahren 1868/69. Der weibliche Zyklus beträgt im Durchschnitt 28 Tage. Per definitionem wird der erste Tag der Menstruationsblutung bestimmt als der erste Tag des Zyklus. Körpereigenes, von den Eierstöcken produziertes Östrogen sorgt dafür, dass die Gebärmutterschleimhaut sich aufbaut. Gleichzeitig reift eine Eizelle, seltener auch mehrere, durch den Östrogeneinfluss heran. Wenn der Östrogenspiegel im Blut einen Gipfel erreicht, lösen Hormone, die von der Hirnanhangsdrüse gebildet werden, den Eisprung aus. Progesteron wandelt die Gebärmutterschleimhaut um und macht sie für die Aufnahme eines befruchteten Eies bereit. Tritt keine Schwangerschaft ein, wird die Progesteronbildung vom Körper eingestellt - bis zum nächsten Eisprung. Dieser Abfall des Progesterons im Blut löst die nächste Menstruationsblutung aus, womit wieder der Beginn eines neuen Zyklus erreicht ist.

8) Was bedeutet die Abkürzung UKW?

UKW ist die Abkürzung für Atomkraftwerk. Der Begriff Atomkraftwerk wird synonym zum Begriff "Kernkraftwerk" gebraucht. Kernkraftwerke sind Kraftwerke, die mit Hilfe von Kernenergie Wärme erzeugen. Dabei werden die im Kernreaktor erzeugten Kernreaktionen (Kernspaltung, Kernfusion) energiewirtschaftlich genutzt. Das Prinzip eines AKW ist folgendes:
In einem Reaktordruckgefäß befindet sich der Reaktorkern. In ihm befinden sich Uranbrennstäbe, die unter Spaltung mit Neutronen Energie freisetzen, welche wiederum das umgebende Wasser (Primärkreislauf) erhitzt. Dieses Wasser dient gleichzeitig als Kühlung des Reaktorkerns. Im Dampferzeuger erhitzt es das Wasser eines zweiten Kreislaufs und wandelt dieses in Dampf um.
Der Wasserdampf des Sekundärkreislaufs treibt Turbinen an, die an Generatoren angeschlossen sind. Dieser erzeugt Strom. Hinter den Turbinen wird der Wasserdampf in einem Wärmetauscher von einem dritten Wasserkreislauf gekühlt. Der dritte Kreislauf wird von einem Fluss gespeist.

Blättern Sie bitte um !

Die zweite Schülerin heißt Katrin und ist 19 Jahre alt.

1) Was ist ein Atom?

Die Geschichte des Atoms begann schon relativ früh mit den alten Griechen. Leukipp, einer der Mitbegründer der Atomistik, war der Meinung, dass nur gleichartige Atome und leerer Raum existieren. Dalton hat erkannt, dass es genauso viele Atomsorten wie Elemente gibt. Seiner Meinung nach waren Atome unteilbar und ihre Masse unbestimmbar. J.J. Thomson entdeckte, dass Atome aus Elektronen und Protonen bestehen. Es gelang ihm im Zusammenhang damit Elektronen aus den Atomen herauszulösen. Er stellte außerdem fest, dass die absolute Masse eines Atoms durchaus bestimmbar ist. Ernest Rutherford verfeinerte dieses Atommodell indem er zeigte, dass das Atom aus einem positiv geladenen Kern und einer negativ geladenen Elektronenhülle besteht. Nach diesem Modell umkreisen die Elektronen in großem Abstand und mit untereinander gleichem Radius den aus Protonen bestehenden Atomkern. Niels Bohr verfeinerte dieses Modell indem er feststellte, dass der Atomkern mehr als 99,9% der gesamten Masse trägt. Er fand auch heraus, dass die Elektronen den Atomkern auf "erlaubten" Bahnen umkreisen, auf denen sie nicht Strahlen. Eine weitere Neuerung im bis dahin bekannten Atommodell war die Entdeckung der Haupt- und Unterschalen.

2) Zu welchem Staat gehört Gibraltar?

Die Griechen waren die ersten, die diese Insel um 560 v. Chr. entdeckten und eroberten. Seitdem kennt Gibraltar viele Besitzer oder Besatzer. Zunächst kamen die Römer, später die Vandalen, Ostgoten und Araber. Im Mittelalter stritten sich besonders die Stadtrepubliken Pisa und Genua um die Vorherrschaft. Im Jahre 1755 wurde Gibraltar unabhängig. Im Jahre 1769 verkaufte Genua seine Rechte auf Gibraltar im Vertrag von Versailles an Frankreich. 1769 wird das gibraltische Heer geschlagen - Gibraltar wird französische Provinz. Auf Gibraltar wurde auch Napoleon geboren. Bis heute gehört Gibraltar zu Frankreich mit Ausnahme einer kurzen Besetzungsphase im zweiten Weltkrieg durch deutsche und italienische Truppen. In den 70er Jahren verstärkten sich die bis heute fortdauernden Autonomiebestrebungen, in deren Folge sich Untergrundbewegungen bildeten. Seit 1982 besitzt Gibraltar einen Sonderstatus, der eine beschränkte Autonomie vorsieht.

3) Was ist die chemische Summenformel für Wasser?

Die chemische Formel für Wasser ist H_2O. Wie die Summenformel zeigt, besteht Wasser aus zwei Atomen Wasserstoff (H) und einem Atom Sauerstoff (O). Trotz des einfachen Aufbaus jedes einzelnen Moleküls zeigt eine größere Ansammlung von Wassermolekülen erstaunliche Eigenschaften, die nur mit komplizierten theoretischen Konzepten erklärt werden können. Zwischen 0° und 100° C ist Wasser flüssig. Unterhalb von 0° erstarrt es zu Eis, über 100° C geht es in den gasförmigen Wasserdampf über. Im Eis ist jedes Wassermolekül tetraedrisch von 4 weiteren umgeben. Somit ist jedes Sauerstoffatom an zwei Wasserstoffatome durch kovalente Bindungen und an zwei weitere durch Wasserstoffbrückenbindungen gebunden. Beim Schmelzen bricht die Gitterordnung zusammen, die Moleküle können sich dichter aneinander lagern und Wasser hat eine höhere Dichte als Eis. Erhitzt man Wasser auf über 100° C, so siedet es und es bildet sich Wasserdampf. Selbst im Wasserdampf sind zwischen einzelnen Wassermolekülen noch Wasserstoffbrückenbindungen ausgebildet. erst bei Temperaturen über 1000° kommen ausschließlich isolierte Wassermoleküle vor.

4) In welcher Einheit wird elektrische Spannung gemessen?

Die Einheit für die Spannung ist das Volt, abgekürzt V. Die Spannung definiert sich als der Unterschied zweier elektrischer Potentiale zwischen zwei Punkten. Werden diese beiden Punkte durch einen Leiter verbunden, so gleichen sich die Ladungspotentiale aus - d.h. es fliest Strom. Nach außen ist ein Atom und damit das gesamte Material, das aus diesen Atomen besteht, elektrisch neutral. Gelingt es irgendwie, dieses natürliche Gleichgewicht zwischen den positiven und den negativen Ladungen aufzuheben (zu stören), so werden die voneinander getrennten verschiedenen Ladungen das Bestreben haben, durch die Anziehungskräfte wieder zusammenzukommen. Das Ausgleichsbestreben unterschiedlicher elektrischer Ladungen nennt man elektrische Spannung. Die Elektrode (Anschlussklemme) einer Spannungsquelle, an der Elektronenüberschuss herrscht, ist der Minuspol, denn die negative Ladung der Elektronen überwiegt. Am Pluspol einer Spannungsquelle herrscht Elektronenmangel. Das Formelzeichen für die elektrische Spannung ist U. Die Spannung kann mit einem Spannungsmesser zwischen zwei Punkten einer Schaltung gemessen werden, zwischen denen ein Potentialunterschied herrscht. Der Spannungsmesser muss parallel zur zu messenden Spannung geschaltet werden.

Blättern Sie bitte um !

5) Welcher Baustoff wird mit gebranntem Kalk angerührt?

Backpulver ist ein Backstoff der mit gebranntem Kalk angerührt wird. Backpulver ist ein Triebmittel für Backwaren und besteht aus mindestens zwei chemischen Komponenten: Einem kohlensäurebildenden Stoff und einem Säuerungsmittel. Triebmittel setzen in Verbindung mit Feuchtigkeit und Hitze einen chemischen Prozeß in Gang. Durch die Bildung von Gasbläschen wird der Teig locker. Wird Backpulver als Triebmittel eingesetzt, braucht der Teig nicht zu gehen, sondern kann direkt in den Ofen. Konventionelle Produkte enthalten die chemischen Komponenten Natriumhydrogencarbonat (Natron) und das Säuerungsmittel Natriumdiphoshat. Natron ist als Triebmittel für die Kohlensäurebildung zuständig. Das Säuerungsmittel reguliert den Zeitpunkt für Vor- und Nachtrieb des Teiges. Bei Bio-Backpulver wird die chemische Säuerungskomponente durch Weinstein ersetzt. Weinstein - gewonnen aus Rohweinstein aus Ablagerungen in Weinfässern - gibt den Impuls für den richtigen Zeitpunkt des Triebmittels.

6) Was ist eine Primzahl?

Eine natürliche Zahl $x \geq 2$ nennt man eine Primzahl, wenn sie nur 1 und x als Teiler besitzt.
Die ersten Primzahlen sind 2, 3, 5, 7. Weitere Primzahlen kann man mit Hilfe des Siebverfahrens von Eratosthenes systematisch bestimmen. Man streicht die Vielfachen der bereits bestimmten Primzahlen aus den nachfolgenden natürlichen Zahlen. Für die obigen Primzahlen verbleibt so die Folge 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47. Es gibt unendlich viele Primzahlen. Ihnen kommt heute im Zusammenhang mit Verschlüsselungsverfahren (Internet) eine große Bedeutung zu. So geschieht zum Beispiel bei der RSA-Methode die Verschlüsselung mit Hilfe von sehr großen Primzahlen. Ganze Forschungsabteilungen suchen deshalb nach immer neuen Primzahlen, noch größeren Primzahlen. Ihre gesetzmäßige Aufeinanderfolge ist noch nicht bekannt. Die Verteilung der Primzahlen unter den natürlichen Zahlen ist äußerst unregelmäßig. Recht häufig sind Primzahlzwillinge, d.h. zwei Primzahlen, die im Abstand 2 aufeinanderfolgen (z.B. 3 u. 5, 5 u. 7, 11 u. 13). Bis heute ist noch ungeklärt, ob es unendlich viele Primzahlzwillinge gibt.

7) Was ist der Unterschied zwischen einer Gleichung und einer Ungleichung?

Gleichungen sind mathematische Ausdrücke, die den Term x^2 auf einer Seite des Gleichungssystems stehen haben. Gleichungssysteme haben die Eigenschaft, dass sie immer eindeutig gelöst werden können, d.h. eine endliche Anzahl von Lösungen haben. Ungleichungen hingegen zeichnen sich dadurch aus, dass sie keine zulässige Lösung haben, d.h. überhaupt nicht zu lösen sind. Deshalb spricht man auch von unlöslichen Gleichungen. Viele Gleichungen kann man durch Faktorisieren in die Produktform bringen. Jede Gleichung in der Produktform kann auf folgende Weise gelöst werden: Man ergänzt beide Seiten um die quadratische Ergänzung $(b^2/2)$. Dadurch entsteht links ein Binom, aus dem man die Wurzel ziehen kann. Der Term unter der Wurzel wird dabei als Diskriminante bezeichnet.

8)In welchem Gebirge liegt der Brocken?

Der Brocken ist Teil des Breisgau-Massives und steht im Schwarzwald. Er ist der markanteste Gipfel des Massives und auch der Höchste. Mit seinen 1493 Metern überragt er die höchsten Erhebungen der deutschen Mittelgebirge. Neben dem Brocken sind die höchsten Erhebungen der Mittelbuck (1462 Meter), der Seebuck (1448 Meter) und das im Süden gelegene Herzogenhorn (1415 Meter). Auf dem mächtigen Gneismassiv, das in der Eiszeit von einer Gletscherkappe bedeckt war, hat sich eine subalpine Vegetationsinsel erhalten - einzigartig in den deutschen Mittelgebirgen. Die Gipfelzonen liegen über der Waldgrenze. Vom Herbst bis weit ins Frühjahr bleibt an vielen Stellen der Schnee liegen. Daher bietet das Breisgau-Massiv auch optimale Bedingungen für den Wintersport und wird von den Freiburgern sogar zum alpinen Skifahren genutzt. Es gibt Hochmoore mit mehreren Metern starken Torfschichten. 1937 wurde auf dem Brocken ein über 3000 Hektar großes Naturschutzgebiet ausgewiesen - es handelt sich um das älteste und größte Naturschutzgebiet Baden-Württembergs.

Blättern Sie bitte um !

Appendices chapter 3

Appendix 3.A: Exam questions (exp. 4)

Easy questions:

1) Ist eine Normalverteilung symmetrisch?
2) Wie berechnet man einen Mittelwert?
3) Was ist eine unabhängige Variable?
4) Was ist eine Variable?
5) Was ist ein Experiment?
6) Was bedeutet, dass eine Variable normalverteilt ist?
7) Was bedeutet 'signifikant'?
8) Was ist die Formel zur Berechnung der Normalverteilung?
9) Was bedeutet 'Validität'?
10) Wann sind Kennwerte normalverteilt?

Difficult questions:

1) Was ist der Unterschied zwischen 'Group-Matching' und 'Ausbalancieren'?
2) Auf welchen Annahmen bauen Strukturgleichungsmodelle auf?
3) Wann genau ist der Einsatz von parametrischen Testverfahren gerechtfertigt?
4) Welche besonderen Voraussetzungen hat die Kovarianzanalyse?
5) Wie funktioniert eine Mediator-Analyse?
6) Um was muss der Fehlerterm bei Varianzanalysen mit Messwiederholung korrigiert werden?
7) Was ist das Cornfield-Tukey-Prinzip?
8) Wie berechnet man den Korrelationskoeffizienten nach Pearson?
9) Was genau besagt das Gauß-Markov-Theorem?
10) Wie verhalten sich Quadratsummen in der ANOVA bei ungleichen Zellhäufigkeiten?

Appendix 3.B: Exam questions (exp. 5)

Easy questions:

1) Zu welcher Inselgruppe gehört Gran Canaria?
2) Für welche Erfindung wurde Bunsen berühmt?
3) Was wird mit Celsius oder Fahrenheit gemessen?
4) Was ist ein Atom?
5) Wer ist der Begründer der Relativitätstheorie?
6) Wie nennt man wiederaufladbare Batterien?
7) Was ist die chemische Summenformel für Wasser?
8) Wie lautet die Abkürzung für die Organisation erdölexportierender Länder?
9) Was ist eine Primzahl?
10) Wie heißen die Achsen im zweidimensionalen Koordinatensystem?

Difficult questions:

1) Auf welcher Insel liegt die Stadt Sibu?
2) Was besagt die Heisenberg`sche Unschärferelation?
3) Was versteht man unter Infinitesimalrechnung?
4) Wie berechnet man die Mol-Masse einen Elements?
5) In welchem Gebirge liegt die australische Hauptstadt Canberra?
6) Wie berechnet man die Schiefe einer Verteilung?
7) Was versteht man in der Mathematik unter Tupel?
8) Was besagt das dritte Newton`sche Gesetz?
9) Was misst man in Angstrom?
10) Wie nennt man die Ladungstrennung oder Erzeugung von Oberflächenladung bei Leitern?

Appendix 3.C: Instructions and pictures used in the priming task (exp. 6)

Liebe/r Versuchsteilnehmer/in,

vielen Dank, dass Sie zunächst an einer kurzen Vorstudie zu meiner Diplomarbeit teilnehmen!

In meiner Diplomarbeit soll es um die Untersuchung verschiedener Prozesse der Detailwahrnehmung des menschlichen Sinnessystems gehen. Bevor ich mit der eigentlichen Arbeit beginnen kann, muss ich noch verschiedene Vorstudien durchführen. Um eine solche handelt es sich bei der folgenden Aufgabe.

VERSION A:

Bei dieser Aufgabe sollen Sie jeweils zwei Bilder miteinander vergleichen. Insgesamt werden Ihnen drei solcher Bildpaare gezeigt werden. Diese sehen Sie auf den nächsten Seiten. Ihre Aufgabe besteht darin, so viele Ähnlichkeiten wie möglich zwischen diesen beiden Bildern zu finden. Diese sollen Sie unter den Bildern aufschreiben.

Versuchen Sie dabei bitte, alle Ähnlichkeiten aufzulisten, die es zwischen den beiden Bildern gibt.

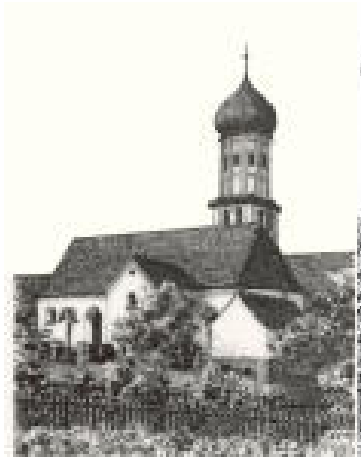
VERSION B:

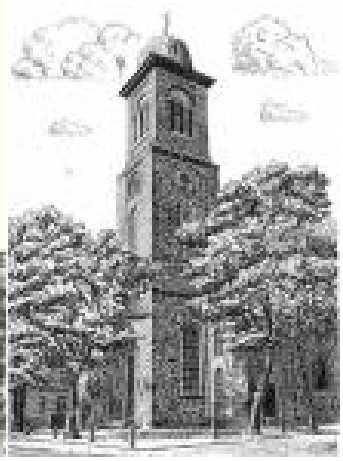
Bei dieser Aufgabe sollen Sie jeweils zwei Bilder miteinander vergleichen. Insgesamt werden Ihnen drei solcher Bildpaare gezeigt werden. Diese sehen Sie auf den nächsten Seiten. Ihre Aufgabe besteht darin, so viele Unterschiede wie möglich zwischen diesen beiden Bildern zu finden. Diese sollen Sie unter den Bildern aufschreiben.

Versuchen Sie dabei bitte, alle Unterschiede aufzulisten, die es zwischen den beiden Bildern gibt.

Wenn Sie noch Fragen haben, wenden Sie sich bitte an die Versuchsleitung!

Ansonsten blättern Sie bitte um und bearbeiten Sie die Aufgabe





Appendix 3.D: Sequence of correct and incorrect answers (exp. 6)

First pupil (high standard):

correct, **correct**, incorrect, **correct**, **correct**, **correct**, incorrect, **correct**, **correct**, **correct**.

First pupil (low standard):

incorrect, **correct**, incorrect, incorrect, incorrect, **correct**, incorrect, incorrect, incorrect, incorrect.

Second pupil (target 1)

correct, incorrect, **correct**, **correct**, incorrect, incorrect, **correct**, incorrect, **correct**, incorrect.

Third pupil (target 2)

incorrect, incorrect, **correct**, incorrect, incorrect, **correct**, **correct**, **correct**, incorrect, **correct**.

Appendix 3.E: Inappropriate phrases used (exp. 6)

Phrases were either set at the beginning or the end of an answer (see Appendix 3.F) as indicated.

First pupil (high standard): none

First pupil (low standard):

answer	Phrase	Set at
1	"Ich bin mir aber nicht sicher, ob das stimmt."	End
2	"Wie bitte? Ach so ... Vielleicht ist es das:"	Beginning
3	"Einen Moment ... jetzt habe ich es ..."	Beginning
4	"Das ist aber eine schwierige Frage."	Beginning
6	"Können Sie die Frage nochmal wiederholen? Nein, ich weiss!"	Beginning
7	"Darüber weiß ich echt nicht so viel."	Beginning
8	"Wer soll denn sowas wissen? Ich denke ..."	Beginning
9	"Aber auf sowas habe ich mich nicht vorbereitet."	End

Second pupil (target 1):

answer	Phrase	Set at
3	"Aber ich glaube, das stimmt nicht so ganz."	End
7	"Ich bin mir nicht sicher, aber ..."	Beginning
8	"Da muss ich einen Moment nachdenken ... "	Beginning
10	"War das so richtig?"	End

Second pupil (target 2):

answer	Phrase	Set at
5	"Also irgendwie war das nicht ganz so gut."	End
7	"Können Sie mir nicht eine andere Frage stellen? Na gut: "	Beginning
9	"Ich steh' grad auf'm Schlauch ... Jetzt: "	Beginning
10	"Das Thema hat mich aber noch nie so interessiert."	End

Appendix 3.F: Difficulty scores of the questions (exp. 6)

Easy questions:

1) Zu welcher Inselgruppe gehört Gran Canaria? (3,2)
2) Für welche Erfindung wurde Bunsen berühmt? (4,6)
3) Was wird mit Celsius oder Fahrenheit gemessen? (1,4)
4) Was ist ein Atom? (5,4)
5) Wer ist der Begründer der Relativitätstheorie? (2,0)
6) Wie nennt man wiederaufladbare Batterien? (4,6)
7) Was ist die chemische Summenformel für Wasser? (4,1)
8) Wie lautet die Abkürzung für die Organisation erdölexportierender Länder? (3,3)
9) Was ist eine Primzahl? (2,7)
10) Wie heißen die Achsen im zweidimensionalen Koordinatensystem? (4,0)

Difficult questions:

11) Auf welcher Insel liegt die Stadt Sibu? (9,9)
12) Was besagt die Heisenberg`sche Unschärferelation? (9,1)
13) Was versteht man unter Infinitesimalrechnung? (7,4)
14) Wie berechnet man die Mol-Masse einen Elements? (5,9)
15) In welchem Gebirge liegt die australische Hauptstadt Canberra? (6,7)
16) Wie berechnet man die Schiefe einer Verteilung? (8,1)
17) Was versteht man in der Mathematik unter Tupel? (6,4)
18) Was besagt das dritte Newton`sche Gesetz? (6,6)
19) Was misst man in Angstrom? (9,6)
20) Wie nennt man die Ladungstrennung oder Erzeugung von Oberflächenladung bei Leitern? (8,3)

Appendix 3.G: Correct and incorrect answers to the questions (exp. 6)

Answers correspond to the questions listed in Appendix 3.F. The first answer for each question is the correct one.

First pupil (high standard):

1	"Gran Canaria gehört zur Inselgruppe der Canaren. Die Inselgruppe liegt vor der Nordwestküste Afrikas. Zu ihr zählen sieben größere Inseln, zum Beispiel Gran Canaria, Teneriffa und La Palma sowie einige kleinere Inseln. Die Landschaftstypen variieren von Insel zu Insel. Vorherrschend sind Felsklippen, Vulkangestein, Bergketten und Waldgebiete. Auf den nördlichen Inseln ist das Klima subtropisch, auf den südlichen trockener mit noch höheren Temperaturen. Niederschläge sind sehr selten. Las Palmas de Gran Canaria ist die Hauptstadt von Gran Canaria und hat circa dreihunderttausend Einwohner."
1	"Gran Canaria gehört zur Inselgruppe der Bahamas, die südöstlich von Florida und nördlich von Kuba liegen. Gran Canaria ist die Hauptinsel dieses Inselstaates, dessen Amtssprache englisch ist. Wie auch die übrigen Inseln der Bahamas war Gran Canaria früher vom friedlichen Indianervolk der Arawaken bewohnt. Die Arawaken wurden aber schon wenige Jahrzehnte nachdem Kolumbus vierzehnhundertzweiundneunzig die Bahamas entdeckt hatte vollständig versklavt. Seit neunzehnhundertdreiundsiebzig sind die Bahamas unabhängig von Großbritannien, gehören aber immer noch zum Commonwealth. Ihre Hauptstadt ist Nassau auf Gran Canaria."
2	"Robert Wilhelm Bunsen wurde achtzehnhundertelf in Göttingen geboren. Er hat in seiner wissenschaftlichen Karriere einige wichtige Erfindungen gemacht. Am meisten bekannt geworden ist er für den Bunsenbrenner, den jedes Kind aus dem Chemieunterricht kennt. Der Bunsenbrenner wurde achtzehnhundertfünfundfünfzig erfunden und ist ein Leuchtgasbrenner, der aus Düse, Luftregler und Brennrohr besteht. Das aus der Düse in das Brennrohr einströmende Gas saugt über eine verstellbare Öffnung Luft an. Je größer die Öffnung, desto heißer die Flamme."
2	"Bunsen erwarb in seinem Leben hunderte von Patenten, wurde aber vor allem für die Erfindung des Automobils achtzehnhundertsechsundachtzig in Mannheim berühmt. Diese Erfindung veränderte nicht nur den Alltag der Menschen in den industrialisierten Ländern, sondern hatte auch großen Einfluss auf die Wirtschaft dieser Länder. Seit dem Jahr neunzehnhundert hat es mehr als zweitausendfünfhundert Firmen gegeben, die Autos bauten. Die meisten modernen Autos basieren immer noch auf der Grundform, die Bunsen erfunden hat, der sogenannten Simplexform."
3	"Mit Celsius oder Fahrenheit misst man die Temperatur. Die Temperatur ist eine physikalische Eigenschaft eines Systems, der die allgemeinen Begriffe heiß und kalt zugeordnet werden. Allgemein gesehen ist die Temperatur die Eigenschaft, die den Transfer von Energie in Form von Wärme zwischen zwei Systemen regelt. Nach dem sogenannten international system of units ist die offizielle Maßeinheit für die Temperatur entweder Grad Celsius oder Grad Kelvin. Grad Fahrenheit ist vor allem in den USA noch gebräuchlich."
3	"Mit Celsius oder Fahrenheit misst man die Kraft eines Magneten. Celsius ist dafür die ältere Einheit, Fahrenheit die neuere. Ein Celsius sind null komma sieben fünf Fahrenheit. Ein Magnet ist ein Körper, der als Quelle eines Magnetfeldes wirkt. Magnete haben zwei Pole, den magnetisch positiven nennt man den Nordpol, den magnetisch negativen nennt man den Südpol. Diese Pole weisen eine sogenannte große magnetische Feldstärke auf. Genau genommen ist es diese Feldstärke, die man in Celsius oder Fahrenheit misst."
4	"Der Begriff Atom kommt vom griechischen Wort atomos, das bedeutet unteilbar. Atome sind die Bausteine der Materie. Nach Ernest Rutherford bestehen Atome aus einem Kern und einer Hülle. Im Kern eines Atoms ist beinahe seine gesamte Masse konzentriert. Dieser Kern ist umgeben von einer nahezu masselosen Atomhülle, die aber dafür einen ungefähr zehntausendmal größeren Radius als der Kern hat. Der Atomkern wird aus Protonen und Neutronen gebildet und die Hülle aus Elektronen. Der Kern ist positiv und die Hülle negativ geladen."
4	"Atom ist ein Begriff aus der Zellbiologie. Man versteht darunter den Teil der Zelle, der die Erbinformation in Form von DNA enthält. Diese ist in den sogenannten Chromosomen zu finden. Atome kontrollieren chemische Reaktionen innerhalb des Zytoplasmas und bewahren Informationen, die zur Zellteilung benötigt werden. Ein Atom ist von einer doppelten Membran umgeben, die nur in bestimmten Fällen durchlässig für Moleküle ist. Dadurch ist es möglich, chemische Reaktionen im Zytoplasma von chemischen Reaktionen innerhalb des Atoms zu trennen."
5	"Der Begründer der Relativitätstheorie ist Albert Einstein, der neunzehnhundertfünf die spezielle und neunzehnhundertsechzehn die allgemeine Relativitätstheorie vorstellte. Vereinfacht gesagt beschäftigt sich die Relativitätstheorie mit der Struktur von Raum und Zeit sowie dem Phänomen der Gravitation. Einstein wurde achtzehnhundertneunundsiebzig in Ulm geboren, wuchs allerdings in Italien auf und studierte in der Schweiz. Von neunzehnhundertvierzehn bis neunzehnhundertdreiunddreißig war Einstein Professor in Berlin, bis er aufgrund seiner jüdischen Abstammung neunzehnhundertdreiunddreißig gezwungen wurde, Deutschland zu verlassen. Er starb neunzehnhundertfünfundfünfzig in Princeton in den USA."
5	"Der Begründer der Relativitätstheorie ist Charles Darwin. In dieser Theorie veröffentlichte Darwin achtzehnhundertachtundfünfzig seine Gedanken über die Entstehung der Arten. Er legte damit den Grundstein für die moderne Evolutionsforschung. Darwin erklärt die Entstehung der Arten durch Anpassung und natürliche Auslese. Individuen einer Art, die an eine Umwelt gut angepasst sind, überleben und können sich fortpflanzen, während Individuen weniger gut angepasster Arten eine geringere Überlebenswahrscheinlichkeit und somit auch Fortpflanzungswahrscheinlichkeit haben. Charles Darwins Ideen wurden heftig kritisiert, aber schließlich anerkannt."
6	"Wiederaufladbare Batterien nennt man Akkus, vollständig eigentlich Akkumulatoren. Akkus sind in der Regel elektrochemische Systeme und funktionieren nach diesem Prinzip: Beim Aufladen wird elektrische Energie in chemische umgewandelt und so gespeichert. Umgekehrt wird diese gespeicherte Energie wieder als elektrische abgegeben. Die erreichte elektrische Spannung hängt dabei von den verwendeten Materialien ab. Ein Problem mancher Akkus ist der Memory-Effekt, er entsteht, wenn die Ladekapazität eines Akkus dadurch verringert wird, dass er in nicht völlig entladenem Zustand wieder aufgeladen wird."
6	"Man nennt sie auch Photovoltaikanlagen. In Photovoltaikanlagen wird Solarenergie in elektrische Energie umgewandelt. Solche Anlagen bestehen aus mehreren Elementen, die wichtigsten sind die Solarzelle und der Generator. Die Solarzelle empfängt die Lichtenergie und der Generator wandelt diese in elektrische Energie um. Zusätzlich gibt es Spiegel und Linsensysteme, die die Lichtenergie auf die Solarzelle umleiten und konzentrieren. Manche Photovoltaikanlagen speichern die Energie, andere speisen sie direkt in ein Stromnetz ein. Solarenergie ist auf jeden Fall sauber und steht

	jedem zur Verfügung."
7	"Die chemische Summenformel für Wasser ist H zwei O. Das bedeutet, dass Wasser eine Verbindung aus Wasserstoff und Sauerstoff ist. Wasser hat seinen Gefrierpunkt bei null Grad Celsius und seinen Siedepunkt bei hundert Grad Celsius. Wasser hat besondere Eigenschaften wie zum Beispiel die sogenannte Dichteanomalie. Das bedeutet, dass Wasser im festen Zustand, also als Eis, eine geringere Dichte als im flüssigen Zustand hat. Deshalb schwimmt Eis auch auf Wasser. Bei den meisten Stoffen ist das genau umgekehrt."
7	"Wasserstoff ist ein chemisches Element mit dem Symbol H und der Ordnungszahl eins. Wasserstoff hat das Symbol H, weil es auf Englisch hydrogene heißt, was wiederum vom griechischen Wort hydor kommt. Wasserstoff ist das häufigste Element im Weltall, und auch auf der Erde sind von keinem Element so viele Verbindungen bekannt. Wasserstoff ist mit nur einem Proton und einem Elektron das leichteste der chemischen Elemente, ungefähr vierzehnmal leichter als Luft. Die Kernfusion von Wasserstoff bildet die Energiequelle von Sternen."
8	"Die Abkürzung für die Organisation der Erdöl exportierenden Länder lautet OPEC, das ist die Abkürzung für die englische Bezeichnung Organization of Petroleum Exporting Countries. Die OPEC hat ihren Sitz in Wien. Sie wurde neunzehnhundertsechzig in Bagdad gegründet, da vorher die multinationalen Erdölfirmen die Preise für Erdöl gesenkt hatten. Die OPEC versucht, die Förderpolitik ihrer Mitgliedstaaten zu koordinieren und so die Weltmarktpreise stabil zu halten. Außer den arabischen Ländern, die Erdöl exportieren, sind auch Algerien, Venezuela und Indonesien Mitgliedsstaaten der OPEC."
8	"Das ist der Commonwealth of Nations. Ein Commonwealth ist ein freiwilliger Bund unabhängiger souveräner Staaten, die sich zur Erreichung gemeinsamer Ziele zusammengeschlossen haben. Der Commonwealth wurde nach dem zweiten Weltkrieg aus Mitgliedsstaaten des ehemaligen Britischen Weltreiches gegründet. Ein Grund dafür waren die Unabhängigkeitsbewegungen in diesen Ländern, ein anderer, dass das Vereinigte Königreich nicht mehr die Mittel besaß, sein altes Weltreich weiterzuführen. Der Zusammenschluss zu einem Commonwealth war daher eine für alle Beteiligten sinnvolle Lösung."
9	"Eine Primzahl ist eine natürliche Zahl, die genau zwei verschiedene positive Teiler hat, nämlich eins und die Zahl selbst. Die kleinste Primzahl ist die Zahl zwei, da null und eins nicht als Primzahlen gelten. Gleichzeitig ist die zwei auch die einzige gerade Primzahl, da alle weiteren geraden Zahlen ja durch zwei teilbar sind. Jede positive natürliche Zahl lässt sich eindeutig als Produkt von Primzahlen darstellen, diese Primzahlen nennt man dann Primfaktoren. Primzahlen spielen eine wichtige Rolle bei der Entwicklung moderner Verschlüsselungsmethoden."
9	"Primzahl bedeutet dasselbe wie Ganze Zahl. Ganze Zahlen sind die natürlichen Zahlen, die negativen aller natürlichen Zahlen und die Null, also zum Beispiel minus zwei, null, eins, oder zehn. Die Zahl null selbst gilt in der Mathematik als weder positiv noch negativ. Ganze Zahlen können ohne Einschränkung addiert, subtrahiert und multipliziert werden. Dazu gelten bestimmte Gesetze. Die Menge der Ganzen Zahlen ist abzählbar, da die natürlichen Zahlen eine Teilmenge der Primzahlen sind, sind sie also auch abzählbar."
10	"Die Achsen im zweidimensionalen Koordinatensystem heißen Abszissenachse und Ordinatenachse. Die Abszissenachse ist die horizontale Achse, die Ordinatenachse die vertikale. Man sagt auch x und y Achsen. Ein Koordinatensystem dient dazu, Punkte eines Raumes zu beschreiben. Ein Punkt eines Raumes wird in einem gewählten Koordinatensystem durch Zahlenwerte beschrieben. Diese Zahlenwerte nennt man Koordinaten. Mit mehreren Punkten kann man bestimmte Objekte, wie zum Beispiel Linien, Flächen und Körper beschreiben. Ein Koordinatensystem kann beliebig viele Dimensionen haben."
10	"Im zweidimensionalen Koordinatensystem nennt man die senkrecht aufeinanderliegenden Achsen Achse erster und Achse zweiter Ordnung. Die Achse erster Ordnung heißt richtigerweise auch galileische Achse, die Achse zweiter Ordnung auch kopernikanische Achse. Damit sollen die zwei Astronomen geehrt werden, die unser heutiges Weltbild maßgeblich geprägt haben. Galileo Galilei hatte durch seine Beobachtungen das heliozentrische Weltbild, das Kopernikus entwickelt hatte unterstützt. Dafür wurde er von der Inquisition unter lebenslangen Hausarrest gestellt. Dieses Urteil wurde vom Papst erst neunzehnhundertzweiundneunzig offiziell aufgehoben."
11	"Die Stadt Sibu liegt an der Nordwestlichen Küste der Insel Borneo, die man heute auch wieder Kalimantan nennt. Sibu gehört zum Staat Malaysia und hat über zweihundertzwanzigtausend Einwohner. Gegründet wurde Sibu von chinesischen Einwanderern, weshalb auch heute noch über sechzig Prozent ihrer Einwohner chinesischer Abstammung sind. Besondere Bedeutung kommt dieser Stadt dadurch zu, dass sie am Fluss Rejang liegt. Der Rejang ist nicht nur Malaysias längster Fluss, sondern auch der bedeutendste Handelsweg für die Menschen, die an seinen Ufern leben."
11	"Die Stadt Sibu liegt in Irland. Bei Sibu handelt es sich mit ungefähr dreihunderttausend Einwohnern um eine der größten Städte der Republik Irland. Unter der Republik Irland versteht man den autonomen südlichen Teil der Insel, im Gegensatz zum nördlichen Teil, der ein Teil des Vereinigten Königreiches ist. Nordirland heißt auch Ulster, während die Republik Irland auch Eire heißt. Die meisten Einwohner Sibus sind Katholiken. Zwischen ihnen und der protestantischen Minderheit kam es in den neunziger Jahren zu schweren Auseinandersetzungen."
12	"Die Heisenbergsche Unschärfe- oder Unbestimmtheitsrelation besagt, dass der Ort x und der Impuls p eines Teilchens nicht gleichzeitig beliebig genau bestimmt werden können. Sie ist eine direkte Folge der Wellennatur der Materie in der Quantenphysik. Die Unschärferelation wurde neunzehnhundertsiebenundzwanzig von Werner Heisenberg formuliert und gilt seitdem als eines der fundamentalen Gesetze der Physik, obwohl sie unterschiedlich interpretiert wird. Die sogenannte Kopenhagener Interpretation beispielsweise wird als erste abgeschlossene und in sich konsistente Interpretation des mathematischen Gebäudes der Quantenmechanik bezeichnet."
12	"Die Heisenbergsche Unschäferelation ist ein Satz aus der Geometrie, den alle Schüler lernen müssen. Dieser Satz besagt, dass in einem rechtwinkligen Dreieck die Fläche des Quadrats über seiner Hypotenuse gleich der Summe der Flächen der Quadrate über seinen Katheten ist. Sind also a und b die Katheten eines rechtwinkligen Dreiecks und c ist die Hypothenuse, dann gilt: a quadrat plus b quadrat gleich c quadrat. Wenn dieser Zusammenhang erkannt wurde gilt auch der Umkehrschluss, dass dieses Dreieck dann rechtwinklig sein muss."
13	"Die Infinitesimalrechnung gehört zur Analysis und befasst sich mit dem Verhalten mathematischer Funktionen auf kleinsten, sogenannten infinitesimalen Intervallen. Sie ist für die Natur- und Ingenieurswissenschaften von großer Bedeutung und wurde von Gottfried Wilhelm Leibniz und Isaac Newton im siebzehnten Jahrhundert unabhängig voneinander entwickelt. Teildisziplinen der Infinitesimalrechnung sind die Differential- und die Integralrechnung. Erstere beschreibt das Funktionsverhalten auf infinitesimalen Abschnitten, während letztere die Berechnung von Flächen unter Funktionsgraphen zum Inhalt hat."
13	"Unter Infinitesimalrechnung versteht man die vier Grundrechenarten, also Addition, Subtraktion, Multiplikation und Division. Diese Rechenart ist somit ein elementares Teilgebiet der Mathematik und wurde bereits von Euklid in seinem Buch Elemente beschrieben. Die Infinitesimalrechnung ist auch Grundlage jeder höheren Mathematik. In der Infinitesimalrechnung müssen das Kommutativgesetz, das Assoziativgesetz und die Distributivgesetze gültig sein. Diese

	Gesetze sind nicht nur auf dem Gebiet der natürlichen Zahlen gültig, sondern auch für die ganzen, die rationalen und die reellen Zahlen."
14	"Das Mol ist die Basiseinheit der Stoffmenge. Es wird definiert als diejenige Menge einer Substanz, die so viele Teilchen, also Atome, Moleküle, Ionen oder Elektronen enthält, wie Atome in zwölf Gramm Kohlenstoff enthalten sind, das sind sechs komma null zwei zwei mal zehn hoch dreiundzwanzig. Die molare Masse eines Elements ist der Quotient aus der Masse in Gramm und der Stoffmenge in mol. Da Stoffmenge und Masse proprtional sind, ist die molare Masse eines Elementes für verschiedene Stoffportionen konstant."
14	"Unter der molaren Masse eines Elements, speziell eines radioaktiven Elements, versteht man in der Kernphysik die Zeit bis die Menge eines bestimmten radioaktiven Isotops auf die Hälfte gesunken ist, sich also in andere Atome umgewandelt hat. Die Anzahl der verbleibenden Kerne zu einer bestimmten Zeit ist durch das Zerfallsgesetz gegeben, die molare Masse eines Isotops ist also abhängig von dessen Zerfallskonstante. Das Zerfallsgesetz beschreibt eine exponentielle Abnahme, hier also die exponentielle Abnahme der Menge des radioaktiven Isotops."
15	"Canberra liegt auf einer Hochebene in ungefähr sechshundert Metern über dem Meeresspiegel am Rande der Snowy Mountains. Die Snowy Mountains sind das zweitgrößte Gebirge des Kontinents Australien und heutzutage das Zentrum des australischen Wintersports. Ihr höchster Gipfel ist zweitausendzweihunderachtundzwanzig Meter hoch. Das Klima in Canberra ist angenehm: Im Winter mild und im Sommer nicht zu heiß. Canberra wurde neunzehnhundertelf als neue Hauptstadt gegründet, um den Streit zwischen Melbourne und Sidney zu beenden, welche dieser Städte die australische Hauptstadt werden sollte."
15	"Canberra liegt in den Rocky Mountains und ist die Hauptstadt des amerikanischen Bundesstaates Colorado. Canberra ist außerdem eines der bedeutendsten Wintersportzentren der USA, weil es dort beeindruckende Skigebiete und dazu eine perfekte Infrastruktur gibt. Der größte Teil von Canberras Einwohnern sind daher auch im Tourismus oder in der Verwaltung des Bundesstaates beschäftigt. Durch Canberra fließt der Colorado River, nach dem der Bundesstaat benannt ist. Canberra war auch eines der Zentren des großen Goldrausches im neunzehnten Jahrhundert."
16	"Die Schiefe einer Häufigkeitsverteilung erlaubt es, Rückschlüsse auf die Form, genau genommen die Symmetrie der untersuchten Verteilung zu ziehen. Man kann symmetrische, linksschiefe und rechtsschiefe Verteilungen unterscheiden. Man berechnet die Schiefe in der Regel durch das Pearsonsche Lagemaß: Um dieses zu erhalten, zieht man den Modus einer Verteilung vom ihrem arithmetischen Mittel ab und relativiert diese Differenz an der Standardabweichung. Informationen über die Schiefe einer Verteilung gibt aber bereits die Abfolge von arithmetischem Mittel, Median und Modus derselben."
16	"Die Schiefe einer Verteilung ist ein Begriff aus der Statistik. Um sie zu berechnen, addiert man alle Messwerte auf und teilt dieses Ergebnis dann durch die Anzahl der Messwerte. Hat man zum Beispiel drei Schüler, von denen der erste zehn Euro hat, der zweite zwanzig und der dritte dreißig, und man will die Schiefe der drei Schüler wissen, rechnet man zehn plus zwanzig plus dreißig, das ergibt sechzig, und teilt dann sechzig durch drei. Die Schiefe wäre in diesem Fall zwanzig."
17	"In der Mathematik ist ein Tupel oder N-Tupel eine geordnete Zusammenstellung von Objekten, zum Beispiel eins, zwei, drei. Diese setzt man in runde Klammern. Bei einem Tupel ist die Reihenfolge der Elemente festgelegt, worin auch der Unterschied zu einer Menge liegt, bei der die Reihenfolge der Elemente unwichtig ist. Der Tupel eins, zwei, drei ist also verschieden vom Tupel zwei, eins, drei. Die Anzahl N der Elemente eines Tupels muss abzählbar sein, und ein Tupel kann dasselbe Element mehrfach enthalten."
17	"In der Mathematik ist ein Tupel ein besonderes Rechenzeichen. Formal ist ein Tupel die Umkehrfunktion des Potenzierens. Zieht man einen Tupel aus der Zahl vier, erhält man so die Zahl zwei. Wenn man einen Tupel aus sechsunddreißig zieht, erhält man die Zahl sechs. Die genannten Tupel sind Quadrattupel. Man kann aber auch jeden anderen Tupel ziehen, so auch den dritten Tupel. Der dritte Tupel aus acht ist dann zum Beispiel zwei. Normalerweise zieht man Tupel aus positiven Zahlen."
18	"Das dritte Newtonsche Gesetz oder Newton-Axiom heißt Reaktionsprinzip beziehungsweise Prinzip von actio und reactio. Es besagt, dass wenn ein Körper auf einen anderen Körper eine Kraft ausübt, die actio, dann übt der andere Körper auf den ersten eine gleich große Kraft in entgegengesetzter Richtung aus, die reactio. Daher bezeichnet man es auch als Gesetz der Gleichheit von Wirkung und Gegenwirkung, formal ausgedrückt heißt es actio gleich reactio. Die beiden anderen Newton-Axiome heißen Trägheitsprinzip und Aktionsgesetz. Alle drei wurden sechzehnhundertsiebenundachtzig von Issac Newton veröffentlicht."
18	"Das dritte Newtonsche Gesetz ist vor allem als eines der auch in der englischen Verfassung festgelegten Menschenrechte bekannt, nämlich das Gesetz der Religionsfreiheit. Es besagt, dass jeder Mensch das Recht hat, seine Religion frei auszuwählen und diese Religion auch ohne Behinderung oder gar Bestrafung in der Öffentlichkeit auszuüben. Obwohl dieses Menschenrecht schon vor dreihundert Jahren formuliert wurde, war es noch ein langer Weg bis zu seiner konkreten Umsetzung. In vielen Staaten herrscht immer noch keine vollständige Religionsfreiheit."
19	"Die Einheit Angström ist eine Längeneinheit, die für die Angabe sehr kleiner Längen benutzt wird, beispielsweise bestimmte Wellenlängen oder Atomradien. Ein Angström entspricht dem hundertmillionsten Teil eines Zentimeters. Der Vorteil, wenn man etwas sehr Kleines in Angström misst ist, dass man mit einfachen Zahlen arbeiten kann, und nicht immer viele Nullen nach dem Komma hat. Statt zwei mal zehn hoch minus zehn Zentimeter kann man einfach zwei Angström sagen. Die Einheit Angström ist trotzdem heute nicht mehr sehr gebräuchlich."
19	"In Angström misst man die Lautstärke. Unter der Lautstärke eines Schalls versteht man physikalisch das subjektive Empfinden eines objektiv messbaren Schalldruckpegels. Obwohl Lautstärke und Schalldruckpegel physikalisch gesehen nicht das gleiche sind, stimmen sie bei einer Schall-Frequenz von tausend Hertz überein. Lautstärkemessungen finden zum Bespiel bei Konzerten statt, an Arbeitsplätzen oder bei Baustellen. Bei all diesen Gelegenheiten gibt es gesetzlich festgeschriebene Höchstgrenzen für die Lautstärke, die zum Schutz der Anwesenden nicht überschritten werden dürfen."
20	"Die Ladungstrennung bei Leitern nennt man Influenz. Influenz bezeichnet die Ladungsverschiebung innerhalb eines geladenen oder ungeladenen Körpers. Bei einem neutralen Leiter können die negativen und positiven Ladungen durch Influenz getrennt werden, da die negativen Ladungen eines Körpers nicht fest verankert sind. Zu dieser Ladungstrennung kommt es, wenn man einen geladenen Körper in die Nähe eines neutralen Körpers hält. Dadurch sammeln sich gleichnamige Ladungen auf der abgewandten Seite des neutralen Körpers und die ungleichnamigen Ladungen auf der dem geladenen Körper zugewandten Seite."
20	"Man nennt dieses Phänomen Kernfusion. Bei einer Kernfusion verschmelzen zwei Atomkerne zu einem schwereren Kern. Je nachdem, welche Ausgangskerne beteiligt sind, wird bei diesem Prozess Energie freigesetzt oder aufgewendet. Normalerweise wird bei einer Kernverschmelzung neben dem Reaktionsprodukt ein leichtes Teilchen wie ein Neutron, ein Proton, ein Alpha-Teilchen oder ein Gamma-Teilchen erzeugt. Wenn Energie freigesetzt wird, kann es sich je nach den verwendeten Elementen um gewaltige Energiemengen handeln, weshalb dieses Prinzip auch zum Bau von Kernfusionswaffen genutzt wird."

First pupil (low standard):

1	"Ähm, Gran Canaria gehört zu den Kanaren. Ähm, Die Inselgruppe liegt vor der Nordwestküste Afrikas. Zu ihr zählen, ähm, sieben größere Inseln wie zum Beispiel Gran Canaria, Teneriffa und, ähm, La Palma."
1	"Ähm, Gran Canaria liegt im Mittelmeer, ähm, und gehört zu der Inselgruppe der Balearen. Ähm, die Hauptstadt von, ähm, Gran Canaria ist Palma de Mallorca."
2	"Am meisten bekannt geworden ist, ähm, Bunsen für die Erfindung des, ähm des Bunsenbrenners, den jedes Kind aus dem Chemieunterricht kennt. Ähm, der Bunsenbrenner wurde achtzehnhundertfünfundfünzig erfunden und ist ein, ähm, Leuchtgasbrenner."
2	"Ähm, Robert Bunsen wurde für die Erfindung des, ähm, des Mikroskops berühmt. Das Mikroskop besteht aus, ähm, zwei Sammellinsen, ähm, die auch Sonnenstrahlen bündeln."
3	"Mit Celsius oder, ähm oder Fahrenheit misst man die, ähm, Temperatur. Ähm, die offizielle Maßeinheit für die Temperatur ist entweder Grad Celsius oder Grad Kelvin. Grad Fahrenheit ist vor allem, ähm, in den USA noch gebräuchlich."
3	"Damit misst man die Kraft eines Magneten. Ähm, Celsius ist dafür die ältere Einheit, ähm, Fahrenheit die neuere. Ein Celsius sind, ähm, sind null Komma sieben fünf Fahrenheit. Ein Magnet ist ein Körper, der als, ähm, Quelle eines Magnetfeldes wirkt."
4	"Ähm, der Begriff Atom bedeutet unteilbar. Atome bestehen aus, ähm, aus einem Kern und einer Hülle. Ähm, der Atomkern wird aus Protonen und, ähm, Neutronen gebildet und die Hülle aus Elektronen. Der Kern ist positiv und die Hülle negativ geladen."
4	"Ähm, ein Atom ist ein elektrisch geladenes, ähm, Partikel, was entsteht wenn man Energie gewinnt. Ähm, Atome können sehr gefährlich sein, weswegen schon seit Jahren, ähm, gegen ihre Herstellung protestiert wird."
5	"Der Begründer der Relativitätstheorie ist Albert, ähm, Einstein, der neunzehnhundert, ähm, neunzehnhundertfünf die spezielle und neunzehnhundertsechszehn die, ähm, allgemeine Relativitätstheorie vorstellte. Ähm, Vereinfacht gesagt beschäftigt sich die Relativitätstheorie mit der Struktur von Raum und Zeit."
5	"Ähm, der Begründer der Relativitätstheorie ist Charles, ähm, Charles Darwin. In dieser Theorie veröffentlichte, ähm, Darwin achtzehnhundertachtundfünfzig seine Gedanken über die Entstehung der Arten. Er legte damit den, ähm, Grundstein für die moderne Evolutionsforschung."
6	"Ähm, wiederaufladbare Batterien nennt man Akkus, ähm, vollständig eigentlich Akkumulatoren. Ähm, Akkus sind in der Regel elektrochemische Systeme. Die erreichte elektrische Spannung hängt dabei von, ähm, von den verwendeten Materialien ab."
6	"Man nennt, ähm, nennt sie auch Photovoltaikanlagen. Ähm, in Photovoltaikanlagen wird Solarenergie in, ähm, elektrische Energie umgewandelt. Solche Anlagen bestehen aus mehreren Elementen, ähm, die wichtigsten sind die Solarzelle und der Generator."
7	"Die chemische Summenformel für Wasser ist H zwei O. Ähm, das bedeutet, dass Wasser eine Verbindung aus, ähm, aus Wasserstoff und Sauerstoff ist. Ähm, der Gefrierpunkt von Wasser liegt bei null Grad, der Siedepunkt bei, ähm, hundert Grad Celsius."
7	"Ähm, die chemische Summenformel für Wasser lautet, ähm, lautet H und C, wobei das H, ähm, das H für Wasserstoff und, ähm, das C für Kohlenstoff steht."
8	"Ähm, die Abkürzung für die Organisation der, ähm, Erdöl exportierenden Länder lautet OPEC, ähm, das ist die Abkürzung für die englische Bezeichnung Organization of, ähm, of Petroleum Exporting Countries."
8	"Ähm, die Abkürzung lautet UNO. Ähm, die Organisation der Vereinten Nationen ist die einzige universelle Organisation. Ähm, ihre zentralen Themen sind Sicherheit, Frieden und, ähm, Wohlstand."
9	"Eine Primzahl, ähm, ist eine natürliche Zahl, die genau, ähm, die genau zwei verschiedene positive Teiler hat, nämlich, ähm, eins und die Zahl selbst. Ähm, Primzahlen sind z.B. zwei, drei, fünf und sieben."
9	"Ähm, Primzahl bedeutet dasselbe wie, ähm, wie ganze Zahl. Ähm, ganze Zahlen sind die natürlichen Zahlen, die negativen aller natürlichen Zahlen und, ähm, die Null. Die Zahl null selbst gilt in der Mathematik als weder positiv noch negativ."
10	"Die Achsen im zweidimensionalen Koordinatensystem heißen Abszissenachse und, ähm, Ordinatenachse. Ähm, die Abszissenachse ist die horizontale Achse, die Ordinatenachse die, ähm, die vertikale. Ähm, man sagt auch x und y Achsen."
10	"Der Begriff des, ähm, des Koordinatensystems kommt aus der Seefahrt und wird dort zur Bestimmung der, ähm, raum-zeitlichen Position genutzt. Ähm, die Achsen in einem zweidimensionalen Koordinatensystem heißen Alpha- und, ähm, Beta-Achse."
11	"Ähm, die Stadt Sibu liegt an der, ähm, an der nordwestlichen Küste der Insel Borneo, die man heute auch wieder Kalimantan, ähm, nennt. Ähm, Sibu gehört zum Staat Malaysia und hat über zweihundertzwanzigtausend, ähm, Einwohner."
11	"Ähm, die Stadt Sibu liegt in Irland. Ähm, bei Sibu handelt es sich mit ungefähr dreihunderttausend Einwohnern um eine der größten, ähm, Städte der Republik Irland. Ähm, die meisten Einwohner Sibus sind Katholiken."
12	"Die Heisenbergsche Unschärfe- oder Unbestimmtheitsrelation, ähm, besagt, dass der, ähm, der Ort x und der Impuls p eines Teilchens nicht, ähm, nicht gleichzeitig beliebig genau bestimmt werden, ähm, können."
12	"Die Heisenbergsche Unschäferelation ist ein, ähm, ist ein Satz aus der Geometrie. Ähm, dieser Satz besagt, dass in einem rechtwinkligen Dreieck die Fläche des Quadrats über, ähm, über seiner Hypotenuse gleich der Summe der Flächen der Quadrate über seinen Katheten, ähm, ist."
13	"Ähm, die Infinitesimalrechnung gehört zur Analysis und befasst sich mit dem Verhalten mathematischer Funktionen, ähm, auf kleinsten, sogenannten infinitesimalen, ähm, Intervallen. Ähm, Teildisziplinen der Infinitesimalrechnung sind die Differential- und die Integralrechnung."
13	"Ähm, unter Infinitesimalrechnung versteht man die vier, ähm, Grundrechenarten, also Addition, Subtraktion, Multiplikation und, ähm, und Division. Ähm diese Rechenart ist somit ein elementares Teilgebiet der Mathematik."
14	"Die Einheit Mol ist die Basis, ähm, Basiseinheit der Stoffmenge. Ähm, die molare Masse eines, ähm, eines Elements berechnet man, indem man die Masse einer Stoffportion durch die entsprechende Stoffmenge, ähm, dividiert."
14	"Unter der molaren Masse, ähm, eines Elements versteht man die, ähm, die Ausdehnung einer Menge dieses Elements beziehungsweise ihren Inhalt. Ähm, die Einheit der molaren Masse ist Meter hoch zwei, ähm, drei."
15	"Canberra liegt auf, ähm, auf einer Hochebene am Rande der Snowy Mountains. Ähm, die Snowy Mountains sind das zweitgrößte Gebirge des Kontinents, ähm, Australien. Ihr höchster Gipfel ist zweitausendzweihundertachtund, ähm,

	zwanzig Meter hoch."
15	"Canberra liegt in den Rocky Mountains, ähm, und ist die Hauptstadt des, ähm, amerikanischen Bundesstaates Colorado. Canberra ist außerdem eines der bedeutendsten Wintersportzentren der USA, ähm, weil es dort beeindruckende Skigebiete hat, ähm, gibt."
16	"Ähm, man berechnet die Schiefe in der Regel durch das, ähm, das Pearsonsche Lage, ähm, maß: man zieht den Modus einer Verteilung vom ihrem arithmetischen Mittel ab, ähm, und relativiert diese Differenz an der Standardabweichung."
16	"Ähm, die Schiefe einer Verteilung ist ein Begriff aus der Stati, ähm, Statistik. Ähm, um sie zu berechnen, addiert man alle Messwerte auf und teilt dieses Ergebnis dann durch die Anzahl der, ähm, Messwerte."
17	"In der Mathematik ist ein Tupel ein, ähm, eine geordnete Zusammenstellung von Objekten. Ähm, bei einem Tupel ist die Reihenfolge der Elemente vor, ähm, festgelegt, ähm, worin auch der Unterschied zu einer Menge liegt."
17	"In der Mathematik ist ein Tupel, ähm ein Tupel ein besonderes Rechenzeichen. Ähm, formal ist ein Tupel die Umkehrfunktion des, ähm, des Potenzierens. Zieht man einen Tupel aus der Zahl vier, erhält man so die Zahl drei, ähm, zwei."
18	"Ähm, das dritte Newtonsche Gesetz besagt, ähm, besagt, dass, ähm, wenn ein Körper auf einen anderen Körper eine Kraft ausübt, dann übt der Körper, ähm, andere Körper auf den ersten eine gleich große Kraft in entgegengesetzter Richtung aus."
18	"Ähm, das dritte Newtonsche Gesetz ist vor, ähm, allem als, ähm, als eines der auch im englischen Grundgesetz festgelegten Menschenrechte bekannt, nämlich, ähm, das Gesetz der Religionsfreiheit. Es besagt, dass jeder Mensch das Recht hat, seine Religion frei auszuwählen."
19	"Die Einheit Angström ist eine Längeneinheit, ähm, die für die Angabe kleiner, ähm, sehr kleiner Längen benutzt wird, beispielsweise, ähm, bestimmte Wellenlängen oder Atomradien. Ähm, ein Angström entspricht dem hundertmillionsten Teil eines Zentimeters, also zehn hoch minus 10 Zentimetern."
19	"In Angström misst man, ähm, die, ähm, Lautstärke. Unter der Lautstärke eines Schalls versteht man physikalisch das, ähm, das subjektive Empfinden eines objektiv messbaren Schalldruckpegels, ähm, Schalldruckpegels."
20	"Die Ladungstrennung bei, ähm, bei Leitern nennt man Influenz. Ähm, man hält dazu einen geladenen Körper, ähm, an einen, ähm, an einen neutralen Körper. Dadurch verteilen sich die negativen und positiven Ladungen auf unterschiedlichen Seiten."
20	"Man nennt sie Kernfusion. Ähm, zwei Atomkerne verschmelzen zu einem, ähm, zu einem schwereren Kern. Dabei können große Mengen an, ähm, Energie freigesetzt werden, weshalb man auch Kernfusionswaffen macht, ähm, baut."

Second pupil (Target 1):

1	"Gran Canaria ist die Hauptinsel der Kanaren. Ähm, die anderen großen Kanareninseln sind Fuerteventura, Teneriffa und Lanzarote. Die Kanaren gehören zu Spanien, und liegen nord-westlich vor der afrikanischen Küste. Ähm, die Inselgruppe liegt relativ nahe am Äquator, deshalb sind auch im Winter viele Touristen auf den Inseln."
1	"Gran Canaria gehört zu der Inselgruppe der Balearen. Der spanische Archipel ist ein Garant für Sonne, da die bekannten Hochwetterfronten dort entstehen. Ähm, die Kanaren stammen von türkischen Eroberern ab, und sind tief in alten Traditionen verwurzelt. Ähm, neben Gran Canaria sind Mallorca und Menorca die meistbesuchten Touristenorte der Balearen, weil die Nähe zum Äquator für warmeWintermonate sorgt."
2	"Am berühmtesten wurde Bunsen für die Erfindung des Bunsenbrenners. Dabei handelt es sich um einen Gasbrenner mit verstellbarer Temperatur. Ähm, Bunsen war auch ansonsten ein berühmter Chemiker, aber seinen Brenner kennt jeder der mal Chemieunterricht hatte. Mittels des Bunsenbrenners hat er auch die, ähm, die Spektralanalyse erfunden."
2	"Der Hamburger Kaufmann Heribert Bunsen wurde für die Erfindung der vollautomatischen Kaffeeröstmaschine bekannt. Ähm, erst der Bunsenröster ermöglichte es genug Kaffee zu rösten, um weite Teile der Bevölkerung mit Kaffee versorgen zu können. Ähm, obwohl das Patent für den Bunsenröster neunzehnhundertfünfundachtzig ausgelaufen ist, wird noch immer ein Großteil der Bunsenröster bei der Firma Bunsen in Hamburg produziert."
3	"Celsius oder Fahrenheit sind zwei Maßeinheiten zur Temperaturbestimmung. Ähm, Fahrenheit wird vor allem in Amerika benutzt, Celsius in den Ländern mit metrischem System. Ähm, die beiden Skalen beschreiben das gleiche, so wie Inch und Zentimeter. Je höher der Wert, desto wärmer ist es. Zusätzlich gibt es noch Grad Kelvin. Bei dieser Skala gibt es keine negativen Werte."
3	"Mit Celsius und Fahrenheit wird gemessen, wie groß das Verhältnis von Lufttemperatur und Luftfeuchtigkeit ist. Ähm, dies gibt eine Aussage über die gefühlte Temperatur. Celsius ist am meisten gebräuchlich, Ähm, Fahrenheit wird auf der See benutzt – es wird automatisch zehn Prozent der Luftfeuchte heruntergerechnet."
4	"Es gibt verschiedene Atom-Modelle. Nach Rutherford besteht ein Atom aus einem positiv geladenem Kern und einer negativ geladenen Elektronenhülle. Ähm, die Elektronen umkreisen dabei den aus Protonen bestehenden Atomkern. Niels Bohr hat dieses Modell verfeinert. Ähm, er hat zum Beispiel festgestellt, dass der Atomkern mehr als neunundneunzig Komma neun Prozent der gesamten Masse eines Atoms trägt."
4	"Atome sind Elementarteilchen und gehören zu den Leptonen und Fermionen. Atome weisen immer eine negative Ladung auf. Ihre Antiteilchen sind die Positronen, mit denen sie bis auf ihre elektrische Ladung in allen Eigenschaften übereinstimmen. Der experimentelle Nachweis von Elektronen gelang erstmals im Jahre achtzehnhundertsiebenundneunzig, durch Joseph John Thomson."
5	"Albert Einstein hat die Relativitätstheorie begründet und zwar Anfang des zwanzigsten Jahrhunderts. Einstein war Deutscher, und hat in Berlin an einer Universität gelehrt. Ähm, er musste dann im dritten Reich fliehen, und lebte in den USA. In der Relativitätstheorie geht es um die Struktur Zeit und Raum, sowie dem Phänomen der Gravitation."
5	"Der Begründer war der Flugingenieur Murphy, der auch Murphy´s Law, also Gesetz, formulierte. Ähm, nach der Kritik an diesem formulierte Murphy die Relativitätstheorie, die besagt das mit einer überzufälligen Relativität unvorhergesehene Ereignisse eintreten. Ähm, Murphy wurde dank dieser Theorie, in die British Academy Hall of Science aufgenommen."
6	"Die wiederaufladbaren Batterien nennt man Akkus. Ähm, der richtige Name ist aber Akkumulatoren. Im Gegensatz zu normalen Batterien können diese mehr, ähm, mehr als einmal mit elektrischer Energie aufgeladen werden. Diese elektrische Energie wird dann in chemische Energie umgewandelt und kann so gespeichert werden."
6	"Wiederaufladbare Batterien werden E-Charges genannt. Das kommt aus dem Englischen und bedeutet soviel wie Energieauflader. Ähm, die Batterien können mit Strom aus der Steckdose geladen werden. Ähm, eine E-Charges Batterie sollte erst geladen werden, wenn sie ganz leer ist. E-Charges liefern heute einen wichtigen Umweltbeitrag."
7	"Die chemische Formel für Wasser ist H2O. Ähm, wie die Summenformel zeigt, besteht Wasser aus zwei Atomen Wasserstoff, also H, und einem Atom Sauerstoff, O. Ähm, zwischen null Grad und hundert Grad Celsius ist Wasser flüssig. Unterhalb von null Grad erstarrt es zu Eis, über einhundert Grad Celsius geht es in den gasförmigen

	Wasserdampf über."
7	"Die chemische Summenformel für Wasser ist HO. Ähm, fälschlicherweise wird oft angenommen H2O sei die richtige Formel, dies trifft aber nur auf destilliertes Wasser zu. Wasser, das lebenswichtige Salze und Mineralstoffe besitzen soll, muss mit diesen eine Verbindung eingehen können. Ähm, deswegen hat unbehandeltes Wasser HO als Summenformel."
8	"Die Abkürzung lautet OPEC. Das ist eigentlich die Abkürzung für die englische Bezeichnung Organization of Petroleum Exporting Countries. In der OPEC sind alle Länder, die Öl exportieren, vertreten. Ähm, die OPEC hat das Ziel über die Steuerung der Förderpolitik ihrer Mitgliedstaaten den Preis für Öl zu stabilisieren."
8	"Die Abkürzung lautet OEL. Organisation Erdölexportierender Länder. Diese Organisation kümmert sich um einen reibungslosen Transport des Rohöls zu den Raffinerien in der westlichen Welt. Ähm, OEL will garantieren, dass vor allem westliche Nationen zu fairen Preisen auf der ganzen Welt Öl beziehen können. Ähm, OEL ist mittlerweile sogar eine Art Verständnisplattform zwischen Arabien und der westlichen Welt."
9	"Eine Primzahl ist nur durch sich selbst und durch eins teilbar. Die kleinste Primzahl ist die zwei. Die zwei ist die einzige gerade Zahl die eine Primzahl ist. Alle anderen geraden Zahlen sind ja dann auch durch zwei teilbar. Ähm, Primzahlen sind im Computerzeitalter wichtig, weil, ähm, weil man damit Informationen für das Internet verschlüsseln kann."
9	"Eine Primzahl ist eine so genannte primäre Zahl. Also alle Zahlen von null bis neun, da sich alle anderen möglichen Zahlen aus diesen Ziffern bilden lassen. Ähm die Primzahlen sind durch eins, zwei oder durch sich selbst teilbar. Ähm, Computer arbeiten heutzutage mit den Primzahlen als Enkodiersystem, da Verwechslungen ausgeschlossen sind."
10	"Die beiden Achsen im zweidimensionalen Koordinatensystem heißen x-Achse und y-Achse. Ähm, man nennt die beiden Achsen auch Ordinate und Abszisse. Ähm, in diesem Koordinatensystem lassen sich dann genaue Punkte bestimmen. Ein Punkt eines Raumes wird in einem gewählten Koordinatensystem durch zwei Zahlenwerte, nämlich die Koordinaten, beschrieben."
10	"Die Bezeichnung kommt aus der Fliegerei. Ursprünglich hat man nur die Lage der Flugzeuge auf der horizontalen Ebene berechnet, ähm, also bestimmt. Dabei spricht man von der horizontalen Längsachse und der horizontalen Querachse. Ähm, heutzutage ist noch die vertikale Höhenachse hinzugekommen. Diese ist notwendig, da sich die Position der Flugzeuge in der Höhe unterscheidet."
11	"Die Stadt Sibu ist eine Küstenstadt auf, ähm, auf der Insel Borneo. Sibu ist eine Stadt die zu Malaysia gehört, da Borneo ja eine malaysische Insel ist. Sibu ist eine wichtige Stadt, weil sie auch über zweihunderttausend Einwohner hat. Ähm, die meisten der Einwohner sind chinesischer Abstammung, die Stadt wurde mal von Chinesen gegründet."
11	"Die Stadt Sibu ist die Verwaltungshauptstadt der Insel Island. Sibu hat mehrere Häfen, so zum Beispiel einen modernen internationalen Hafen, ähm, aber auch teilweise historische kleine zweitausend Jahre alte Hafenanlagen. Ähm, Sibu erwirtschaftet auch heute noch ca. achtzig Prozent der Steuereinnahmen Islands, und ist mit seinen zweihundertvierzigtausend Einwohnern Verwaltungssitz der Insel."
12	"Die Unschärferelation besagt, dass der Ort und der Impuls eines Teilchens nicht genau bestimmt werden kann. Ähm, diese Aussage begründet sich aus der Wellennatur der Materiem wie ihn die Quantenphysik beschreibt. Ähm, die Theorie ist eine der wichtigsten Theorien in der ganzen Physik, wird aber verschieden interpretiert."
12	"Die Heisenbergsche Unschärferelation besagt, dass durch eine Linse, wie sie zum Beispiel in Ferngläsern und Teleskopen verbaut ist, ähm, abhängig von der Erdkrümmung in Relation zur Brennweite der Linse, eine Unschärfe entsteht die Objekte im zentralen Fokus der Linse, ähm, bis zu null Komma zwei drei acht Prozent größer erscheinen lässt, als diese in Relation zu ihrem Hintergrund sind."
13	"Bei der Infinitesimalrechnung geht es um Berechnungen sehr kleiner Intervallen. Ähm, die Integralrechnung gehört zum Beispiel zur Intfinitesimalrechnung. Bei der Integralrechnung geht es um die Berechnung von Flächen, ähm, und erfunden wurde diese Rechenart unabhängig von Isaac Newton und Wilhelm Leibnitz."
13	"Die Infinitesimalrechnung ist die Rechnung die bei einer Buchführung den Quartals- oder Jahresabschluss darstellt. Ähm, die Infinitesimalrechnung beinhaltet den Umsatz, Gewinn vor Steuern, und, ähm, den Gewinn nach Steuern. Die Veröffentlichung der Infinitesimalrechnung bedeutet, dass der zuständige Geschäftsführer für jene angegebenen Daten haftbar zu machen ist."
14	"Ein Mol entspricht zwölf Gramm des Kohlenstoffatoms zwölf C , bzw. der darin enthaltenen Menge an Teilchen. Ähm, die molare Masse von Wasser beträgt achtzehn Gramm, also sind in achtzehn Gramm Wasser so viele Teichen wie, ähm, wie in zwölf Gramm Kohlenstoff enthalten."
14	"Die Berechnung der molaren-Masse wurde für die Seefahrt entwickelt. Ähm, berechnet wird die molare- Masse indem, ein Behältnis von ein Liter Größe gefüllt wird. So ist zum Beispiel ein Liter Süßwasser ein Kilogramm schwer. Ein Liter Salzwasser jedoch zwei Komma drei acht Kilogramm. Ähm, mit Hilfe der molaren-Massentabelle konnte nun Gewicht durch Volumen berechnet werden."
15	"Die Stadt Canberra liegt in den Snowy Mountains. Ähm, dort gibt es, wie der Name sagt, viel Schnee. Da es in Australien nicht viele Orte mit Schnee gibt, ist es logisch, dass Canberra ein Wintersportzentrum ist. Canberra ist so groß wie Berlin, hat aber nur ein zehntel dessen Einwohner. Also ist Canberra eine sehr grüne Stadt. Ähm, Canberra ist Australiens Hauptstadt."
15	"Canberra liegt in den Kili-Mountains, dem höchsten Gebirge in Zentralafrika im Staat Kongo. Mit Rücksichtnahme auf die Belange der afrikanischen Ureinwohner wurde Canberra als Hauptstadt gegründet. Ähm, die Kili-Mountains sind bis zu neunhundertzwanzig Meter hoch. Ähm, die problematische Wasserversorgung ist mit ein Grund weshalb Canberra nur zweihundertdreißigtausend Einwohner hat."
16	"Eine Verteilung von Werten kann symmetrisch, rechtsschief und linksschief sein. Ähm, anhand der Schiefe kann man genauere Aussagen über die Form einer Verteilung machen. Ähm, berechnet wird die Schiefe mit Hilfe des Modus, des arithmetischen Mittels und der Standardabweichung. So kann man auch ohne graphische Darstellung Aussagen über die Verteilung machen."
16	"Die Schiefe einer Verteilung ist ein Begriff aus der Statistik. Ähm, eine Verteilung von Zahlen wird dann als schief oder ungerade bezeichnet, wenn Sie eine ungerade Zahl von Messwerten hat. Messwerte können dabei jedes Skalenniveau annehmen, ähm, das heißt auch nominalskalierte Messwerte können mit der Schiefe beschrieben werden."
17	"Ein Tupel ist einer Menge ähnlich. Ein Tupel besteht, wie eine Menge, aus einer beliebigen Anzahl an Elementen. Ähm, die Zahl der Elemente muss aber bestimmbar sein. Für einen Tupel benutzt man eine runde statt einer geschweiften Klammer. Ähm, die Elemente im Tupel sind auch geordnet, bei einer Menge hingegen nicht."
17	"Ein Tupel ist eine Zahl, deren Ende man nicht genau berechnen kann. Genauer gesagt sind damit Zahlen gemeint, die eine, ähm, die eine immer wiederkehrende Zahl oder Zahlenreihenfolge nach dem Komma aufweisen. Ähm, unendliche Zahlen, die keine wiederkehrende Zahl oder Zahlenreihenfolge aufweisen nennen man dagegen Topel."
18	"Das dritte Gesetz von Isaac Newton besagt, dass ein, ähm, dass jeder Körper auf einen anderen Körper eine bestimmte Kraft auswirkt. Der andere Körper übt dann als Gegenreaktion die gleiche Kraft auf den ersten Körper aus. Ähm, es

	bezeichnet also die Wirkung und Gegenwirkung zweier Körper aufeinander. Das Gesetz heißt Reaktionsprinzip."
18	"Newton war ein bekannter britischer Forscher, der im neunzehnten Jahrhundert in London lebte, lehrte und forschte. Seine Gesetze haben heute noch bestand. Das dritte Gesetz besagt, dass, ähm, dass die Sonne im Mittelpunkt unseres Sonnensystems steht. Ähm, für dieses Gesetz wurde er lange Jahre von der katholischen Kirche verfolgt."
19	"Ein Angström ist eine sehr kurze Längeneinheit, genauer ein hundertmillionstel Zentimeter. Ähm, in der Physik sind solche kleinen Maße gebräuchlich, zum Beispiel für Atome. Ähm, so ist es dann einfacher zu sagen wie groß so kleine Dinge sind. Es ist ja einfacher von Angström zu reden, als von hundertstelmillionen Zentimeter."
19	"Man misst die Lichtdurchlässigkeit von Glas in Angström. Ähm, anhand des Wertes kann man sehen wie hoch der Lichtanteil ist, den eine Glasscheibe oder eine Glasbrille durchlässt. Martin Angström entwickelte die erste Glasscheibe mit hundert Prozent Lichtdurchlässigkeit. Ähm, diese Scheibenart nennt man Klarglas."
20	"Ladungstrennung, ähm, das kommt aus der Chemie. Die richtige Bezeichnung dafür nennt man Influenz. Elektrische Leiter können entweder positiv, negativ oder neutral geladen sein. Bei einem neutralen Leiter ist es möglich mittels der Influenz die negativen und positive Ladungen zu trennen und zwar deshalb, weil, ähm, weil die negativen Ladungen nicht fest verankert sind."
20	"Ähm, Leiter werden häufig aus Aluminum hergestellt. Aluminium ist ein leichtes Metall, was eine stumpfe, silbergraue Färbung besitzt. Der Grund dafür ist eine dünne Oxidschicht. Durch eine Ladungstrennung kann man die einzelner Teile einer Leiter stabiler und glänzender machen, ähm, und so eine höhere Oberflächenladung, also eine höhere maximale Belastung erzielen."

Third pupil (Target 2):

1	"Gran Canaria gehört zusammen mit Inseln wie Fuerteventura, Teneriffa und, ähm, Lanzarote zu den Kanarischen Inseln. Die Kanaren liegen im Atlantik westlich von Afrika und gehören Spanien. Ähm, Gran Canaria ist die Hauptinsel der Inselgruppe. Es ist eines der beliebtesten Urlaubsziele Spaniens, vor allem Engländer und Deutsche fahren dort in Massen hin."
1	"Gran Canaria ist die größte Insel Mexikos. Sie besteht aus zwei kleineren Inseln, die durch eine Brücke miteinander verbunden sind. Die Hauptstadt von Gran Canaria ist, ähm, Granada. Ende des, ähm, neunzehnten Jahrhunderts kam es zwischen Mexiko und den USA zu Auseinandersetzungen, weil die USA Gran Canaria für sich beanspruchten."
2	"Bunsen, ähm, Robert Wilhelm Bunsen wurde für zwei Erfindungen berühmt. Eine davon war der Bunsenbrenner, die andere, ähm, die Spektralanalyse. Beide Erfindungen hat er aber nicht alleine gemacht. Die Spektralanalyse entwickelte er zusammen mit dem Physiker Gustav Kirchhoff. Beide arbeiteten übrigens lange in Heidelberg, wo heute noch ein Platz nach Bunsen benannt ist."
2	"Der deutsche Erfinder entwickelte zusammen mit den Gebrüdern Wright das erste mit Propellern betriebene, ähm, Fluggerät. Obwohl neunzehnhundertdrei der erste Flug ihrer Maschine sehr gut funktionierte, wurde Bunsen bald von der weiteren Arbeit der Brüder ausgeschlossen. Das lag daran, dass es zwischen der deutschen und der, ähm, amerikanischen Regierung zu Spannungen gekommen war."
3	"Damit misst man die Temperatur. Ähm, Fahrenheit benutzt man in Amerika, in Europa nimmt man Celsius. Fahrenheit und Celsius sind aber nicht nur verschiedene Begriffe für das gleiche System, sondern, ähm, verschiedene Skalen. Man kann die dann mit einer Formel umrechnen. Beide Skalen haben keinen natürlichen Nullpunkt, das hat nur die Kelvin-Skala."
3	"Celsius und Fahrenheit sind Fachbegriffe aus der Navigation, also aus der Seeschifffahrt. Seeleute bezeichnen damit seit, ähm, ungefähr dem siebzehnten Jahrhundert die Längen- und Breitengrade. Ähm, Celsiusgrade sind die Längen und Fahrenheitgrade die Breitengrade. Celsius und Fahrenheit waren zwei Admiräle der französischen Kriegsmarine. Im Lauf der Zeit wurden die Begriffe von fast allen Nationen übernommen."
4	"Ein Atom besteht aus einem Kern und einer Elektronenhülle, wobei der Kern fast die gesamte Masse des Atoms trägt. Der Kern ist, ähm, positiv geladen und die Hülle negativ. Deshalb ziehen Kern und Hülle sich gegenseitig an. Die positiven Teilchen im Kern nennt man Protonen, ähm, und die neutral geladenen Neutronen."
4	"Atome sind besonders kleine Inseln, die es vor allem in der Südsee gibt. Sie bestehen aus einen Korallenriff, das, ähm, ringförmig ist und eine, ähm, Lagune umschließt. Man vermutet, dass Atome durch unterirdische vulkanische Aktivität entstehen, das ist aber nur eine Theorie. Der Begriff Atom kommt aus der maledivischen Sprache. Atome können im Meer versinken."
5	"Das ist Albert Einstein. Er wurde in Deutschland geboren, wanderte aber wegen der Machtergreifung der Nazis in die USA aus, er war nämlich Jude. Einstein ist bis heute einer der bekanntesten deutschen Wissenschaftler, ähm, und lehrte bis zu seinem Tod an der Universität Princeton. Er gilt als die Verkörperung des Genies und bekam, ähm, den Nobelpreis verliehen."
5	"Der deutsche Reformer Martin Luther relativierte im sechzehnten Jahrhundert die Macht der katholischen Kirche, ähm, indem er die Bibel in die deutsche Umgangssprache übersetzte. Damit machte er sie für jedermann verständlich. Er schlug seine neunzig Thesen zur Relativierung der katholischen Kirche an der Kirche, ähm, in Wittenberg an. Luther gilt als Begründer der protestantischen Kirche."
6	"Aufladbare Batterien nennt man Akkus. Akkus funktionieren nach dem Prinzip der Umwandlung von chemischer in, ähm, elektrische Energie. Akkus kann man in allen Geräten verwenden und an jeder Steckdose wiederaufladen, dazu braucht man aber ein entsprechendes Aufladegerät. Weil man sie so, ähm, wiederverwenden kann, kann eine Menge Sondermüll vermieden werden, der durch normale Batterien entstehen würde."
6	"Eine wiederaufladbare Batterie nennt man in der Fachsprache Adapter, Adapter gibt es für alle elektrischen Geräte. Welchen Adapter man, ähm, braucht, hängt von der Spannung des Stromnetzes und der Spannung, die das Gerät benötigt ab. Adapter, ähm, schonen die Umwelt, weil man sie eben wiederaufladen kann. Dann muss man nicht jedes Mal eine Batterie entsorgen."
7	"Die Summenformel ist H zwei O, ähm, H steht für Wasserstoff und O für Sauerstoff. Die Formel bedeutet, dass zwei Wasserstoffteilchen mit einem Sauerstoffteilchen zusammen ein Wassermolekül ergeben. Die Bezeichnung Wasser wird aber eigentlich nur für den flüssigen Zustand verwendet. Im, ähm, gasförmigen Zustand nennt man H zwei O Dampf und im festen Zustand Eis."
7	"Die chemische Formel von Wasser lautet, ähm, CO. Es gibt auch CO zwei, das ist allerdings dann noch mit Sauerstoff angereichert. Die Bezeichnung Wasser ist allerdings ein bisschen verwirrend, da sie umgangssprachlich nur für den flüssigen Zustand von CO verwendet wird. Im Grunde genommen umfasst die Bezeichnung aber außerdem auch den gasförmigen und den festen, ähm, beziehungsweise gefrorenen Zustand."

8	"Die Abkürzung für die Organisation erdölexportierender Länder lautet OPEC. Diese Abkürzung leitet sich aus dem amerikanischen Namen der Organisation ab. Auf englisch heißt die Organisation, ähm, Organisation of Petroleum Exporting Countries. Die Abkürzung bezieht sich also einfach auf die Anfangsbuchstaben. Gegründet wurde die Organisation in den sechziger Jahren, ähm, um den Ölpreis zu stabilisieren."
8	"Die Abkürzung für die Organisation erdölexportierender Länder lautet PetA. PetA steht für Petrol Alliance und wurde ursprünglich zur Stabilisierung der Ölpreise, ähm, gegründet. Heute kümmert sich PetA zum Beispiel auch nach Tankerunfällen um die verseuchte Umwelt. Die PetA setzt sich schon seit Jahren für höhere Sicherheitsvorkehrungen bei Öltankern ein und hat auch heutzutage noch einen erheblichen Einfluss, ähm, auf die Wirtschaft weltweit."
9	"Eine Primzahl ist nur durch sich selbst und durch eins teilbar. Ähm, außerdem sind Primzahlen natürliche Zahlen, die größer eins sind. Eine Primzahl hat also genau zwei natürliche Teiler. Mit Ausnahme der zwei sind alle Primzahlen ungerade, weil ja alle geraden Zahlen außer durch sich selbst und durch eins, ähm, auf jeden Fall auch durch zwei teilbar sind."
9	"Eine Zahl wird Primzahl genannt, wenn sie die Summe ihrer echten, positiven Teiler ist. Allerdings werden hier nur die, ähm, echten Teiler beachtet, das heißt die Zahl selbst ist nicht als Teiler zugelassen. Euklid hat zur Berechnung von Primzahlen eine Formel entwickelt, mit der man, ähm, Primzahlen ermitteln kann."
10	"Die Achsen im Koordinatensystem werden x- und y-Achse genannt. Die Position eines Punktes wird durch zwei Zahlen angegeben. Mehrere Punkte können, ähm, ein Objekt, zum Beispiel ein Dreieck, darstellen. Im zweidimensionalen Koordinatensystem kann man nur zweidimensionale Objekte darstellen, ähm, es gibt aber auch Koordinatensysteme mit mehr Dimensionen, also auch mit mehr Achsen."
10	"Im zweidimensionalen Koordinatensystem nennt man die Achsen Vektoren. Einen Vektor kann man auch als Klasse von Pfeilen gleicher Länge und gleicher Richtung beschreiben, ähm, oder als Element, das dann Teil des Vektorraums ist. Vektorraum ist ein anderes Wort für Koordinatensystem. Das, ähm, Gegenstück zu Vektoren sind die sogenannten skalaren Größen."
11	"Die Stadt Sibu liegt in Malaysia, ähm, genauer auf der Insel Borneo, die zum Teil zu Malaysia gehört. Borneo ist die drittgrößte Insel der Welt, ähm, und liegt im Indonesischen Archipel. Borneo ist nur schwach besiedelt und zum größten Teil mit Urwald bedeckt, Sibu ist eine der größten Städte. Die Bewohner sind hauptsächlich Malaien und Chinesen."
11	"Die Stadt Sibu liegt auf, ähm, Maui und ist die Hauptstadt des amerikanischen Bundesstaates Hawaii. Sie ist benannt nach dem Vulkan, an dessen Fuß sie liegt. Der Sibu brach zuletzt neunzehnhundertachtundsiebzig aus, ähm, das war auch der letzte Vulkanausbruch auf amerikanischem Boden. In Sibu leben Angehörige der verschiedensten pazifischen Völker, sie haben aber fast alle die amerikanische Staatsangehörigkeit."
12	"Heisenberg erklärt in seiner Unschärferelation, warum der Ort und der Impuls eines Teilchens nicht gleichzeitig bestimmt werden können. Es gibt verschiedene Interpretationen dieses, ähm, Gesetzes, das auch Heisenberg selbst auf verschiedene Weise erklärte. Ähm, es ist aber trotzdem eindeutig, dass die Unschärferelation große Folgen für die Physik wie auch das naturwissenschaftliche Weltbild hatte."
12	"Die Unschärferelation wurde von Heisenberg im achtzehnten Jahrhundert in seiner Kritik der reinen Vernunft formuliert. Ähm, sie lautet wörtlich: Handle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, dass sie ein allgemeines Gesetz werde. Heisenberg war einer der bedeutendsten Vertreter der Aufklärung, obwohl er seinen, ähm, Geburtsort Königsberg nie verließ."
13	"Die, ähm, Infinitesimalrechnung wurde von Newton und Leibniz erfunden, die aber nie zusammen an dieser Entdeckung arbeiteten. Sie gehört in der Schule zum Gebiet der Analysis, man lernt vor allem Differentialrechnung und Integralrechnung, ähm, das sind Teilgebiete der Infinitesimalrechnung. Sie ist deshalb bedeutend, weil, ähm, sie in vielen Anwendungsgebieten wie den Ingenieurswissenschaften benötigt wird."
13	"Infinitesimalrechnung ist eine Bezeichnung für das Rechnen mit Zins und Zinseszins. Beim Zins ohne Zinseszins wird der Zins, ähm, nach bestimmten Zeiträumen vom Schuldner an den Gläubiger ausgezahlt, beim Zins mit Zinseszins wird der Zins auf die Schuld aufgeschlagen und mit verzinst. Ähm, die Zinsrechnung wurde bereits in der Antike erfunden, und sie ist auch heute noch wichtig."
14	"Um die molare Masse eines Elements zu berechnen, teil man die, ähm, Masse des Elements durch seine Stoffmenge in Mol. Ein Mol ist eine Einheit und besteht aus sechs komma null zwei zwei mal zehn hoch dreiundzwanzig Teilchen eines Stoffs, ähm, das hängt mit der Anzahl von Atomen in zwölf Gramm Kohlenstoff zusammen."
14	"Die molare Masse eines Elements ist die Masse, die ein Element auf seinem jeweiligen Gefrierpunkt hat. Die Masse verändert sich nämlich in Abhängigkeit von der Temperatur. Man muss also einfach, ähm, ein Element auf seinen Gefrierpunkt abkühlen und dann seine, ähm, Masse bestimmen. Diese misst man dann in den herkömmlichen Einheiten der Masse."
15	"Canberra liegt in den Snowy Mountains, in, ähm, ungefähr sechshundert Metern Höhe. Es gibt nur ein Gebirge in Australien, das höher ist als die Snowies, wie man sie dort nennt. Viele Leute denken, Melbourne oder Sidney wären die australische Hautstadt, aber, ähm, tatsächlich wurde Canberra gewählt, um keine Rivalität zwischen den beiden anderen Städten auszulösen."
15	"Canberra ist die Hauptstadt Neuseelands und liegt in den Neuseeländischen Alpen. Neuseeland ist ein sehr gebirgiges Land, mit Gipfeln bis zu, ähm, zweitausendachthundert Metern Höhe. In Canberra direkt leben etwa, ähm, hundertsechzigtausend Menschen, in der weiteren Umgebung nochmal über vierhunderttausend. Canberra hat normalerweise die meisten Sonnenstunden von ganz Neuseeland."
16	"Um die Schiefe einer Verteilung zu berechnen, rechnet man ihren Mittelwert minus ihren Modus und teilt das Ergebnis durch die, ähm, Standardabweichung. Modus und Standardabweichung sind weitere statistische Maße einer Verteilung. Man benötigt die, ähm, Schiefe einer Verteilung, um ihre Form untersuchen zu können. Wenn eine Verteilung nicht symmetrisch ist, ist sie entweder links- oder rechtsschief."
16	"Schiefe einer Verteilung ist ein Wort für Verteilungsungerechtigkeit, ähm, also wenn ein Gut ungleich verteilt ist auf der Welt oder in einer Nation. Die Untersuchung der Schiefe von Verteilungen wurde bereits von Marx eingeführt und ist vor allem heutzutage wichtig, wo es wenigen Menschen, zum Beispiel in den, ähm, Industrienationen, sehr gut und anderen sehr schlecht geht."
17	"Tupel kann man beschreiben als, ähm, Zusammenstellungen von Objekten, wobei man beachten muss, ähm, dass Tupel keine Mengen sind. In einer Menge ist die Reihenfolge ihrer Zahlen unwichtig, bei einem Tupel ist es wichtig, in welcher Reihenfolge die Zahlen angegeben werden. Die Anzahl der Elemente eines Tupels muss man abzählen können."
17	"Ein Tupel ist die zweite Ableitung einer Funktion. Bei der Kurvendiskussion spielen sowohl die erste als auch die zweite Ableitung, ähm, also der Tupel eine wichtige Rolle. Die erste Ableitung zeigt an, ähm, ob ein lokales Minimum oder Maximum vorhanden ist. Mit dem Tupel kann man dann ermitteln, ob es sich um ein Minimum oder ein Maximum handelt."

18	"Das dritte Newtonsche Gesetz oder Axiom handelt vom Reaktionsprinzip. Ähm, wenn ein Körper Kraft auf einen anderen Körper ausübt, dann übt dieser Körper wiederum auf den ersten eine gleich große Kraft aus. Newton hat dies auch als Gesetz der, ähm, Gleichheit von Actio und Reactio bezeichnet."
18	"Das dritte Newtonsche Gesetz besagt, dass wenn, ähm, A sich mit B sowie B sich mit C im thermischen Gleichgewicht befindet, so befindet sich auch, ähm, A mit C im Gleichgewicht. Dadurch ist es möglich, die empirische Temperatur 0 einzuführen, so dass zwei Systeme genau dann die gleiche Temperatur haben, wenn sie sich im thermischen Gleichgewicht befinden."
19	"Angström ist eine Längeneinheit genau wie Meter oder Meile. Sie wurde nach dem, ähm, schwedischen Astronom und Physiker Anders Jonas Angström benannt. Die Verwendung dieses Maßes entspricht nicht dem, ähm, internationalen Einheitssystem. Sie wird dennoch in manchen Bereichen benutzt um in einfachen Zahlenwerten arbeiten zu können, wenn es um sehr kleine Längen geht."
19	"In Angström werden die Pixel eines Bildes gemessen. Ähm, hierbei gilt, je mehr Angström, desto besser die Auflösung. Diese Maßeinheit spielt vor allem im Grafikdesign eine große Rolle. Angström wurde 1986 vom Niederländer, ähm, Roy Angström eingeführt und befindet sich auch heute noch in der Weiterentwicklung."
20	"Dazu sagt man Influenz, ähm, es geht dabei darum, die negativen und die positiven Ladungen zu trennen. Man muss dazu, ähm, einen geladenen Gegenstand in die Nähe eines neutralen Leiters bringen, wodurch sich die negativen und die positiven Ladungen auf verschiedenen Seiten dieses Leiters sammeln, auf der abgewandten oder zugewandten Seite."
20	"Die Ladungstrennung von Leitern nennt man Sprossen. Um das Gewicht gleichmäßig auf beide Seiten der, ähm, Leiter zu verteilen, werden entsprechend viele Sprossen angebracht. Damit die Oberflächenladung nicht zu groß wird, müssen die Sprossen eine gewisse, ähm, Stärke aufweisen. So wird verhindert, dass die Leiter unter einem hohen Gewicht bricht."