ELSEVIER

Learning and Instruction 16 (2006) 401–415

Learning and Instruction

www.elsevier.com/locate/learninstruc

The role of scales in student ratings

Peter Sedlmeier*

Chemnitz University of Technology, Department of Psychology, 09107 Chemnitz, Germany

Abstract

Questions used in course evaluations should mainly measure the quality of teaching, and students' answers to those questions should not be influenced by other factors. This paper investigates how seemingly neutral rating scales and multiple-choice questions might have an impact on the results of such evaluations. In several studies, it has been shown that the way in which a scale is constructed may strongly influence the answers it elicits in surveys and tests. Whether and to what extent common course-related ratings of students are also affected by the kinds of scales used is the main topic of this paper. Four studies examined the influence of scale polarity (unipolar vs. bipolar scales), the role of different ranges in time scales, and the impact of ordering choices in a certain way. Participants' ratings and answers were strongly influenced by all these manipulations. It is recommended to pay special attention to the role of scales when constructing questionnaires for course evaluations and when interpreting course evaluation reports.

Keywords: Scales; Student ratings; Bias; Course evaluation

Course evaluations play an ever-increasing role in all kinds of educational settings. However, there is some evidence that students' judgments might be influenced by factors other than the quality of teaching. For instance, apart from the kind of course (lectures vs. seminars), the amount of interest students have in a given course seems to have a strong impact on their overall ratings, thus lowering ratings for required courses compared to ratings for courses that can be chosen freely (Spiel & Gössler, 2000). Another factor that should not play a role but apparently does is grading leniency (Greenwald & Gillmore, 1997a). Also, more formal topics such as research methods and statistics seem to receive systematically lower ratings (Diehl, 2001). Moreover, student ratings have been found to be influenced by several characteristics of instructors (e.g., rank, experience, autonomy, and personality traits), by course discipline, course level, and even by the gender of teachers and students (Basow, 1995; d'Appolonia & Abrami, 1997; Murray, Rushton, & Paunonen, 1990). Very importantly, evaluation results are also influenced by the kind of criteria used: even seemingly equal criteria can yield vastly different results (Lebherz, Mohr, Henning, & Sedlmeier, 2005).

Despite these findings, the majority of the researchers who work in the field nonetheless support the use of student evaluations to improve teaching. However, this recommendation usually refers not to ad hoc constructed

* Tel.: +49 371 531 36431; fax: +49 371 531 27419.
E-mail address: peter.sedlmeier@phil.tu-chemnitz.de

0959-4752/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.learninstruc.2006.09.002

questionnaires but to standardized tests that also provide norms (e.g., Cashin & Downey, 1992; Marsh, 1982; Rindermann, 2003; Rindermann & Amelang, 1994; Staufenbiel, 2000). The influence of potential moderator variables is taken care of in three different ways by supporters of course evaluations. First, the influence of a given factor might be interpreted as indicative of teaching quality (e.g., heightened interest in a given course might be an indicator of the high opinion students have about the teacher of that course). Second, the impact of the factors mentioned above might be judged to be not or not so influential after all: for instance, a re-analysis of the evidence about the impact of grading leniency on students' ratings indicates that the effects are much smaller than initially assumed (e.g., Marsh & Roche, 2000). Third, some authors propose applying weighting and correction procedures on the raw data to arrive at unbiased ratings (e.g., Diehl, 2003; Greenwald & Gillmore, 1997b; Marsh, 1995).

If all the potential threats to the validity of student ratings discussed above can be dealt with appropriately, can we be confident that we get unbiased ratings? This paper argues that this might not always be the case: seemingly innocent differences in the way the scales themselves are constructed might have a remarkable influence on the results. To date, there is no agreed-upon standard for how scales for course ratings should be constructed and there is considerable variation in what the scales of published questionnaires look like (e.g., Abrami & d'Appolonia, 1990; Diehl, 2001). Moreover, many attempts to evaluate some aspect of teaching quality seem to rely on ad hoc questionnaires with scales that are also constructed ad hoc.

Questions in course ratings commonly ask about several aspects of course content and structure as well as the instructor's behavior and personality characteristics. Answers to these questions usually require students to mark a value on a scale or to choose from a set of response alternatives. The present research did not address whole evaluations but focused on commonly used aspects of course evaluation reports that also have been discussed as potentially exerting a strong impact on course ratings: interest in the topic and perceived relevance of a course (e.g., Cashin & Downey, 1992; Spiel & Gössler, 2000) and workload (e.g., Greenwald & Gillmore, 1997a; but see Marsh & Roche, 2000).

If it turned out that variations in the way scales are constructed influence student ratings on course-related topics, this could constitute potential threats to the validity of course evaluation reports. So it seems important to find out whether such biases exist, and if so, how strong they are. Even small but systematic biases could have a detrimental influence on the correct interpretation of results in students' evaluation of teaching.

Studies 1, 2a, and 2b examined the impact of differently constructed scales on ratings about these aspects. Studies 2a and 2b as well as Study 3 were also concerned with judgments about evaluation criteria. Before the studies are described in more detail, it is necessary to discuss why and how scales can influence judgments.

1. Pragmatic implications of scales

When we try to understand a piece of text or listen to somebody, we do not only rely on linguistic rules but also guard social rules. According to Grice (1975), we obey *conversational maxims* governed by the cooperative principle: ''Make your conversational contribution such as required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged'' (p. 67). This means, that we expect conversational contributions to be as informative as necessary, relevant, truthful, brief, orderly, and unambiguous. If the semantic meaning of a sentence is not fully specified (as it very seldom is), we draw what Harris and Monaco (1978) termed a pragmatic implication. To draw that pragmatic implication, we might use whatever information is available, including information from our memory. If we are not certain about the contents of our memories, the wording of a sentence we read or a question we hear may serve to draw pragmatic implications about what we remember. For instance, when Loftus (1975) asked participants ''Do you get headaches *frequently*, and, if so, how often?'' the reported number of headaches was about three times as high as when participants were asked ''Do you get headaches *occasionally*, and, if so, how often?''

In many studies, Schwarz and his collaborators have shown that pragmatic implications are not only drawn in everyday conversations but also by respondents in research situations and surveys (for summary see Schwarz, 1999). Respondents' answers can be influenced by the context in which a question is embedded, as for instance, by information about what the researcher might be possibly interested in or information conveyed

in adjacent questions. Another possibility to influence answers is to construct response scales in a certain way. Seemingly irrelevant details of such scales may have a strong impact on the responses. For instance, when asked to rate how successful they had been in life, 13% of respondents given an 11-point rating scale that ranged from −5 (not at all successful) to +5 (extremely successful) endorsed a value between −5 and 0. When, however, the numeric values were changed to range from 0 (not at all successful) to 10 (extremely successful) the proportion of participants who answered in the lower half of the scale (0–5) more than doubled to 34% (Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991). Apparently, the bipolar scale (−5 to +5) had participants interpret the statement ''not at all successful'' differently than the unipolar scale. Whereas the pairing of ''not at all successful'' with ''0'' might lead to an interpretation of ''absence of success,'' the pairing of ''not at all successful'' with ''−5'' might mean ''presence of failure.'' In another typical study, Schwarz, Hippler, Deutsch, and Strack (1985) asked about daily TV consumption. To one group of respondents, they provided a collection of response alternatives ranging from ''up to ½ h'' to ''more than 2½ h'' in increments of ½ h. Only 16.2% of the respondents indicated that they watched more than 2½ h per day. Another group received a scale beginning with ''up to 2½ h'' and continuing in steps of ½ h to ''more than 4½ h.'' In this group 37.5% of the respondents indicated that they watched more than 2½ h TV per day. According to Schwarz et al. (1985), respondents assume that the middle range of the scale represents the ''average'' or ''usual'' behavior: when ''more than 2½ h'' was at one extreme of the scale (first group), this category was less likely to be chosen than when it covered the middle range (second group).

The scales discussed so far can be considered interval or ordinal scales. However, pragmatic implications also seem to play a role with nominal scales. A very subtle effect of scale construction with nominal scales was observed by Attali and Bar-Hillel (2003). In several studies, they found that when dealing with multiple-choice tasks, both test makers and test takers show a preference for the middle positions. For instance, they asked faculty members of a psychology department, of whom most had experience in writing multiple-choice tests for their students, to write a four-choice question on any topic they wished; they obtained a distribution of correct answers over the four positions (7, 21, 30, 7) that clearly favored the middle positions. Attali and Bar-Hillel (2003) observed the same tendency on the test takers' side. In one study, psychology undergraduates were to imagine that they were taking a multiple-choice test with four options. One question was: ''What is the capital of Norway?'' Participants were not given any options but only the positions A, B, C, and D as choices. This was done to ensure that only guessing played a role. Again the frequency distribution for the four positions was heavily middle centered (7, 33, 28, 1). This tendency to favor middle positions is also evident in other kinds of tasks, for instance, ''hide and seek tasks'' and apparently, people are generally not aware of it (see Attali & Bar-Hillel, 2003; Bar-Hillel & Attali, 2002, for overview).

In sum, the way scales are constructed has been shown to systematically affect the answers in survey research and even in standardized tests in many ways. This paper explores whether and to what extent also students are affected by pragmatic implications possibly contained in questions as used in the context of course evaluations.

2. Study 1

Study 1 aimed to explore whether the way scale endpoints are numbered influences judgments about important aspects of course evaluation. In particular, it was hypothesized that participants use the whole range in unipolar scales, that is, scales that contain only positive numbers (e.g., 0–10), whereas in bipolar scales (e.g., −5 to +5), participants were expected to concentrate on the negative (e.g., −5 to 0) or the positive part (e.g., 0–5) depending on whether the issue to be judged is generally regarded as negative or positive, respectively. The study also examined whether the influence of different endpoint numberings was less strong for issues about which participants can be assumed to be relatively sure than for issues they were probably not so certain about. Participants were probably quite certain about the *interest* they had in their main subject of study, because they might have chosen that subject after a long decision process. In contrast, they could be expected to be less sure about the *relevance* of the course in which they were enrolled because the contents of this course were new for them. If scales serve informative functions, the information contained in the way the scales are numbered should be used to a greater extent in the second case.

2.1. Method

2.1.1. Participants

Sixty-seven students at the Chemnitz University of Technology participated in the study in a classroom setting as part of a course requirement. All participants studied for a minor in psychology and the course in question was on research methods in psychology. Most of the students were enrolled in educational science degree programs.

2.1.2. Materials and procedure

In the first session of the course, after they had been thoroughly informed about the course content, participants filled in a one-page questionnaire about their expectations of the course (original texts in German). There were two versions of the questionnaire that differed in the endpoint numbering of scales for two of the questions that were situated in the lower half of the questionnaire.

In the *unipolar version*, the first of these questions read "How interesting do you find your (main) subject of study? Please give a number between 0 = totally uninteresting and 100 = extremely interesting." In the *bipolar version*, the wording was identical except that "0" and "100" were replaced by "−50" and "+50." The second relevant question in the unipolar version read: "How relevant do you think this course (research methods in psychology) is for your subject of study? Please give a number between 0 = irrelevant and 100 = very relevant." Again, the bipolar version of the question was identical to the unipolar version except that "0" and "100" were replaced by "−50" and "+50."

The questionnaires were distributed from a stack that contained 80 questionnaires with the two versions in alternating order. The stack was divided into several parts that were handed over to participants who were to take one sheet and pass the rest to their neighbors. Of the 67 participants, 31 happened to complete the bipolar and 36 the unipolar version. Surplus exemplars of the questionnaires were collected by the experimenter.

2.2. Results

After participants had completed the questionnaires, they were asked whether they had additional remarks about the questions but none apparently noticed that there were two different questionnaire versions. This was confirmed by participants when the results of the study were discussed in the classroom some weeks later.

Mainly box plots and effect sizes are used to report the results in this and further studies. Box plots give a good overview of both measures of central tendency and variation, and effect sizes are concise and contain more information than significance tests (for why effect sizes are more informative than significance tests see Rosenthal, Rosnow, & Rubin, 2000; Sedlmeier, 1999, Appendix C). Correlational effect sizes were used throughout.[1] In addition, p values − from the results of two-sample t-tests unless stated otherwise − are also reported whenever applicable, for the interested reader. Because effect sizes as well as significance tests can be strongly biased by outliers, the effect sizes without outliers usually give a more precise picture. In the following, effect sizes both with and without outliers (if applicable) are reported.

Fig. 1 shows box plots for the two questions (here and in all further analyses 50 was added to the bipolar ratings to make the two scale versions comparable). The left pair of box plots reveals a small median difference between participants' ratings of how interested they were in their (main) topic of study depending on the endpoints of the scales. However, the effect size indicates that this difference should not be given too much weight. The correlative effect sizes expressing the difference between the two conditions are $r = -.02$ ($p = .86$) for all values and $r = .04$ ($p = .75$) when the outlier in Fig. 1 (left) is removed from analysis.

[1] Correlational effect size measures that express the difference between two means can either be calculated from the test statistic (mostly t) or as the correlation between group membership (e.g., unipolar = 0, bipolar = 1) and the dependent variable (see Rosenthal et al., 2000). For the general equivalence of correlative effect sizes and effect sizes expressed as standardized differences see Rosenthal (1994) and Sedlmeier (1996).

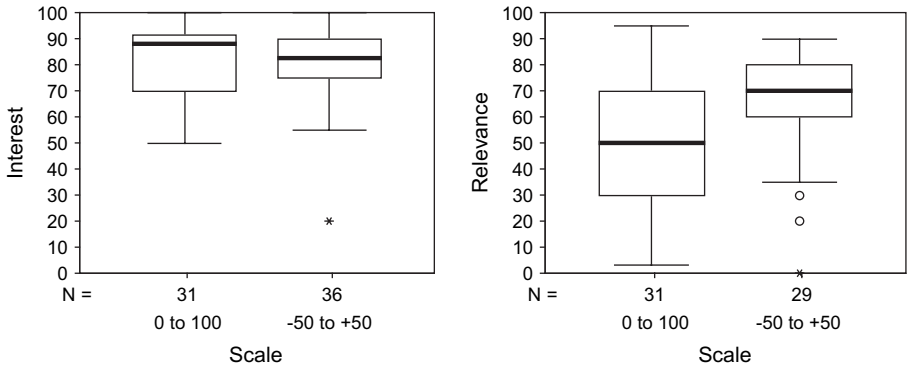

Fig. 1. Box plots showing the results of endpoint numbering in Study 1. The left and right pairs of box plots depict participants' ratings for the interest question and the relevance question, respectively.

In contrast to the ratings about participants' interest in their main subject of study, the judged relevance of the current course for their subject was strongly influenced by the way the scale endpoints were numbered (Fig. 1, right). Seven of the participants in the bipolar condition did not answer this question. The median rating in the bipolar condition was 20 scale points higher than that in the unipolar condition. Expressed as a correlation between condition and ratings this amounts to an $r = .33$ ($p = .011$) for all values and $r = .46$ ($p = .0003$) when the three outliers were removed.

2.3. Discussion

Consistent with the results in survey research, the numbering of scale endpoints had a marked influence on participants' judgments about a central aspect of course evaluations: how relevant they judged the course for their subject. An interpretation for that effect in line with previous explanations (Schwarz, 1999) is that participants' overall attitude about the course was slightly positive, and that, therefore, they tended to use only the positive part of the bipolar scale (0 to +50), whereas they used the whole range of the unipolar scale. Participants may have taken the numbers to disambiguate the meaning of scale values (see Schwarz et al., 1991). This effect was strong, conforming to a large effect according to Cohen's (1992) effect size conventions. The influence of the scale endpoints did not, however, extend to judgments about a topic participants could be expected to be relatively sure of: their interest in their main subject of study. Although there is a difference in level between the ratings for relevance and interest (the higher interest ratings might be in part due to self-worth reasons), this result indicates that degree of certainty may play a moderating role in the effects caused by endpoint labeling: the more certain one is about a topic, the less influence the scale has on ratings.

3. Study 2a

Study 2a examined whether the results found in Study 1 could be replicated. In addition, it served to explore two other kinds of scale effects. First, it pursued the question of whether participants' expectations of course workload were influenced by the labeling of time scales, and second, it explored whether ratings about the importance of course evaluation criteria were influenced by the position (middle or outside) in which they were presented on a questionnaire.

3.1. Method

3.1.1. Participants

Seventy-five students at the Chemnitz University of Technology took part in the study in a classroom setting as part of a course requirement. Participants were from the same pool as participants in Study 1 and took the course on research methods in psychology 1 year afterward. Again, most of the students studied educational science.

3.1.2. Materials and procedure

Materials and procedure for the two questions that examined the impact of endpoint numbering were the same as in Study 1. In addition, the questionnaire included two more critical questions. One question used a time scale similar to the one employed by Schwarz et al. (1985) and asked about the prospective workload: "What is your estimate: how much time will you spend on average to prepare for one session of this course and go over the lesson again afterwards? Please check one of the following." In each of two versions, five choices were offered. In the *small-scale version*, the options were "less than 5 min," "5–15 min," "16–30 min," "31 min up to 1 h," and "more than 1 h." The respective categories in the *large-scale version* were "less than 30 min," "31 min up to 1 h," "1 h up to 1.5 h" "1.5 h up to 2 h," and "more than 2 h."

The other critical question was about the choice of criteria for course evaluation. The question read: "If you were to judge the quality of a course, which of the following four criteria would be the most important for you?" In one version, the criteria were presented in this order (in one line): "exam orientation," "practical orientation," "structuredness," and "difficulty." In the other version, the ordering of the terms was "practical orientation," "exam orientation," "difficulty," and "structuredness." Three of the four criteria used had been rated as very important by Chemnitz students in former evaluation studies (Boehnke, Petri, Baier, & Goldschmidt, 2000) and one ("difficulty") has been found to play an important role in students' evaluations (e.g., Greenwald & Gillmore, 1997a, but see Marsh & Roche, 2000).

The three different types of questions (i.e., polarity, time range, and ordering, two versions for each type of question) were counterbalanced across questionnaires. A supply of 100 randomly ordered questionnaires was divided into several parts that were handed over to participants who were to take one questionnaire and pass the rest to their neighbors. Surplus exemplars were collected by the experimenter.

3.2. Results

As in Study 1, it turned out that the different questionnaire versions went unnoticed by the participants.

3.2.1. Polarity

Fig. 2 shows the results for the impact of endpoint numbering in participants' ratings of their interest in their main subjects of study (left) and the relevance of the course (research methods in psychology) for their subject (right).

Thirty-seven participants had answered the questions in the unipolar version and 38 in the bipolar version. For both questions, ratings were higher with the bipolar scale. For the question about the interest in the subject of study, the median difference was 5 scale points and for the question on the relevance of the course, this difference was 25 scale points. The effect size for the interest question (no outliers) was $r = .12$ ($p = .29$); and the effect sizes for the relevance question were $r = .32$ ($p = .006$) for all cases and $r = .48$ ($p = .000003$) with the outliers (Fig. 1, right) excluded.

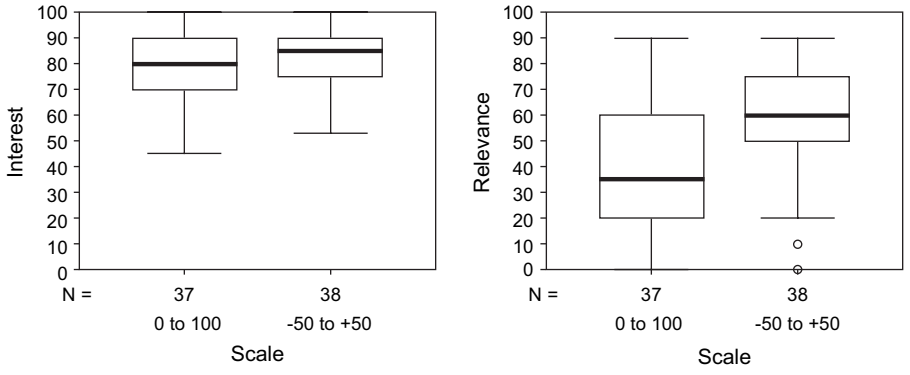

Fig. 2. Box plots showing the results of endpoint numbering in Study 2a. The left and right pairs of box plots depict participants' ratings for the interest question and the relevance question, respectively.

Table 1
Time estimated to be used to prepare for and go over the content of one session of the course in Study 2a

Small scale			Large scale		
Time interval	Number of participants	Percentage of participants	Time interval	Number of participants	Percentage of participants
Less than 5 min	1	2.7	Less than 30 min	3	8.3
5–15 min	2	5.4	31 min up to 1 h	22	61.1
16–30 min	12	32.4	1 h up to 1.5 h	7	19.4
31 min up to 1 h	19	51.4	1.5 h up to 2 h	4	11.1
More than 1 h	3	8.1	More than 2 h	0	.0
Total	37			36	

3.2.2. Time scale

Table 1 reports the number of participants who marked a given time interval in the small-scale and the large-scale versions. The time range offered affected participants' judgments. If one takes the percentage of participants who expected to spend 1 h or more to prepare a session of the course and review the lesson afterward, there is a large difference. In the small-scale condition, only 8.1% planned to do so, whereas in the large-scale condition the respective percentage was almost fourfold: 30.5%. This difference amounts to an $r = .28$ ($p = .015$, χ^2 test).

3.2.3. Ordering of evaluation criteria

For each version, the percentage with which each of the four criteria was marked was determined. It turned out that the four criteria were assigned quite different amounts of importance (Fig. 3). Practical orientation was judged most important followed by structuredness and exam orientation. None of the participants marked difficulty as the most important criterion. A comparison between the percentage with which a given criterion was chosen when it appeared at the outside vs. the middle positions revealed a slight preference for the middle position (Fig. 3). Whereas the summed percentage for the middle positions was 105.5%, it was only 94.5% for the outside positions.

3.3. Discussion

All three types of potential influences of scaling examined in Study 2a, polarity, time range, and ordering, showed an impact on participants' ratings. The effect for the relevance ratings for the course was again strong: participants who used a bipolar scale rated the relevance the course had for their subject of study much higher than those who used a unipolar scale. In contrast to Study 1, there was a small effect in the expected direction of polarity on participants' ratings about how interested they were in their main subject of study. However, this effect was much smaller

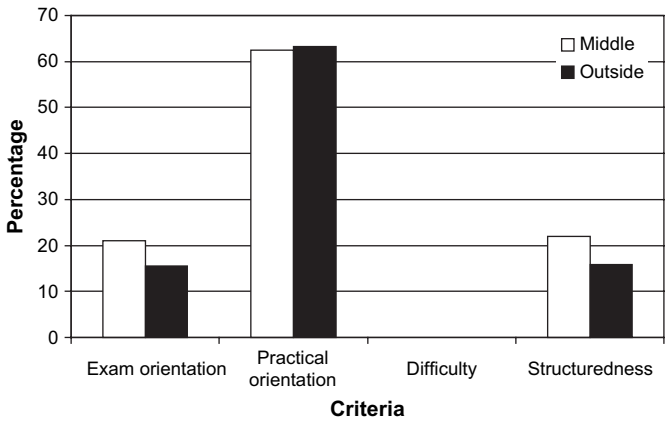

Fig. 3. Percentages with which any of four evaluation criteria were chosen as "most important" in Study 2a. Percentages add up to 200 because each criterion was judged in a middle and an outside position.

than the one found for the relevance question and, consistently with the result in Study 1, indicates that certainty about an issue might lessen the impact of endpoint numbering.

The result for the different time scales was also consistent with those found previously in survey research (Schwarz et al., 1985): participants tended to mark a ''typical'' time interval in the middle of the scale. The effect was strong and yielded large increases in the time participants expected to spend for the course when they were given the large-scale compared to the small-scale version. Finally, the ordering of response alternatives was in accordance with the findings of Attali and Bar-Hillel (2003): course evaluation criteria were preferred when placed in the middle of a list over when they were placed in the outside positions. In the current example, the effect found could, for instance, lead to a preference reversal for exam orientation vs. structuredness, depending on whether these criteria appeared in the middle or the outside positions. However, this effect was small and not totally consistent and therefore a replication seemed to be necessary.

4. Study 2b

Study 2b was a replication of Study 2a. In particular, it sought to find out whether the small effect for the nominal scale (difference between middle and outside positions) could be replicated. It also pursued the question of whether amount of involvement in an issue might moderate the size of the effect induced by differently labeled scales. For participants who are very much involved in a given question, the disambiguating information of scales might not be so necessary to arrive at a judgment (because they also might rely on other kinds of information) and therefore might exert a smaller influence on their ratings. Thus, the overall effects might decrease (as compared to those found in Study 2a) in a group that could be expected to be more involved in the content of the course.

Unlike their counterparts in the educational systems of the United States and some other countries, German psychology students specialize from the very beginning and complete their studies with a diploma that is comparable to a master's degree. There is a high demand for the subject, and there are nationwide space restrictions; therefore psychology students are a highly selected group with high school grades as the main selection criterion. Because it is very difficult to get a place in a psychology program, those students selected can be expected to be more involved in their studies and in the content of the courses they chose than students who study psychology only for a minor with no restriction of admission (like the participants in Studies 1 and 2a). This difference in involvement between psychology students and students whose main subject of study is not psychology is especially evident in methods courses. Thus, if involvement in the course content lessens the impact of the way a scale is constructed, the effects in this study (using psychology students) should be generally smaller than the effects found in Study 2a.

4.1. Method

4.1.1. Participants

Seventy-six psychology (diploma) students at the Chemnitz University of Technology took part in the study in a classroom setting as part of a course requirement. The course content in Study 2b was similar to that in the former studies although several topics were treated a bit more in depth.

4.1.2. Materials and procedure

Similar to the studies reported above, participants received a questionnaire after a thorough introduction into the content of the course. The critical questions dealing with the possible influence of polarity, time scale, and ordering on students' ratings were identical to those in Study 2a.

4.2. Results

As in Studies 1 and 2a, participants did not indicate any awareness about the variations in questionnaires both after the study and some weeks later when the results were disclosed in class.

4.2.1. Polarity

Both the ratings for interest in psychology as well as those for the relevance of the course were markedly higher than in Study 2a (Fig. 4), indicating higher involvement. The high overall level of the responses might have easily led

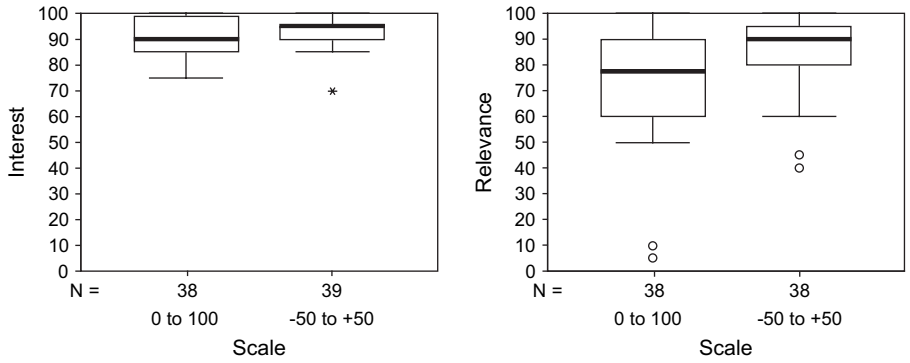

Fig. 4. Box plots showing the results of endpoint numbering in Study 2b. The left and right pairs of box plots depict participants' ratings for the interest question and the relevance question, respectively.

to a ceiling effect but nonetheless, the effect of polarity was still pronounced with a median difference of 5 scale points for the interest question and a median difference of 12.5 scale points for the relevance question (Fig. 4). The respective effect sizes for the interest question were $r = .17$ ($p = .141$) for all students and $r = .23$ ($p = .047$) without the outlier. For the relevance question, the effect sizes were $r = .33$ ($p = .004$) for all participants and $r = .38$ ($p = .001$) without the four outliers (Fig. 4).

4.2.2. Time scale

The impact of different time scales on participants' responses was similar to that in Study 2a. Table 2 shows that again participants tended to use the middle part of the scale more than the outer parts. Whereas 14.6% of the participants in the small-scale condition intended to spend more than 1 h per course session, the respective percentage in the large-scale condition was 27.8%, which amounts to a $r = .16$ ($p = .16$, χ^2 test).

4.2.3. Ordering of evaluation criteria

Fig. 5 shows a pattern similar to that found in Study 2a: practical orientation was rated as the most important criterion for course evaluation. Again, participants tended to assign more importance to the criteria when they appeared in the middle vs. the outside positions. This effect was even more pronounced than in Study 2a. The summed percentages for the middle positions were 111.3% compared to 88.7% for the outside positions.

4.3. Discussion

As expected, the level of the ratings of interest, relevance, and expected time investment was generally higher in Study 2b than in Study 2a, indicating a higher amount of involvement of psychology (diploma) students in the content

Table 2
Time estimated to be used to prepare for and go over the content of one session of the course in Study 2b

Small scale			Large scale		
Time interval	Number of participants	Percentage of participants	Time interval	Number of participants	Percentage of participants
Less than 5 min	3	7.3	Less than 30 min	6	16.7
5–15 min	7	17.1	31 min to 1 h	20	55.6
16–30 min	12	29.3	1 h up to 1.5 h	5	13.9
31 min up to 1 h	13	31.7	1.5 h up to 2 h	2	5.6
More than 1 h	6	14.6	More than 2 h	3	8.3
Total	41			36	

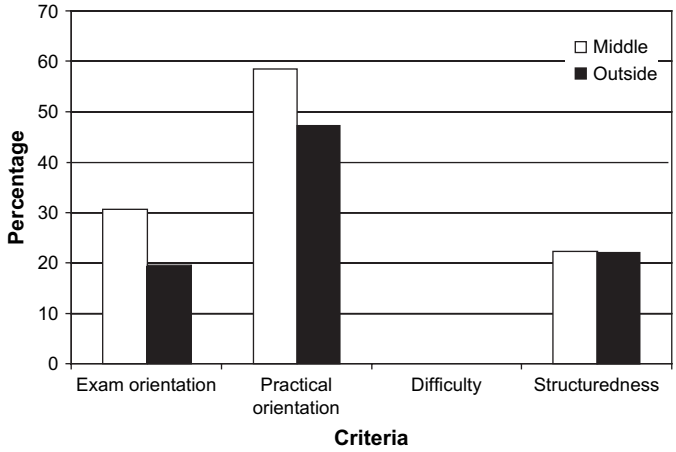

Fig. 5. Percentages with which any of four evaluation criteria were chosen as "most important" in Study 2b.

of the course. Apart from that difference, Study 2b successfully replicated the results found in Study 2a. All three kinds of manipulation had the expected effects on students' ratings. Again, the effect of polarity was smaller for the interest question than for the relevance question, thus confirming the moderating role of participants' certainty about an issue. Despite the difference in level, participants' ratings in this study were not less influenced by the information contained in the scales than in Study 2a. The effect of the ordering of response alternatives was again not very large but it turned out to be quite stable. Again, the present results indicate possible preference reversals between exam orientation and structuredness depending on the position of these two criteria. In sum, it seems that, in contrast to degree of certainty, amount of involvement does not markedly influence the impact of scales.

5. Study 3

A comparison of the results in Studies 2a and 2b suggests that the effect of involvement with the content of the course on students' course-related ratings is small at best. However, it might be that higher involvement with the rating task itself (instead of involvement in the course content) might decrease the influence of different scaling, such as different endpoint numbering. Study 3 explored this question. Instead of rating criteria presented to them, participants in Study 3 were asked to create evaluation criteria themselves and then rate the importance of those criteria. Generating criteria themselves instead of just making judgments about those criteria should make participants think more about the issue and, therefore, should lead to a higher involvement in the rating task. This, in turn, might make the effect of endpoint numbering vanish.

5.1. Method

5.1.1. Participants

Twenty-seven psychology students who took part in a course on evaluation research served as participants in this study. Data were collected as part of a classroom activity. None of the participants had taken part in Study 2b.

5.1.2. Materials and procedure

Participants were handed a sheet of paper on which they first had to fill in some demographic information. Then they read "Please list those criteria that you would use for course evaluation. Please indicate for each criterion how important you judge it to be." Below that, the sheet contained two columns, one headed "criterion" and the other "importance." For 14 of the participants, the instruction indicated the importance of the criterion by using a number between 0 = unimportant and 100 = very important. For the other 13 participants, "0" and "100" were replaced by "−50" and "+50." The sheet contained 10 lines where participants could fill in a criterion and the corresponding importance rating. The procedure for distributing the sheets was the same as in the other studies.

Table 3
Criteria listed, number of persons who listed them, and average importance ratings produced in Study 3

Criteria for course evaluation	Unipolar scale (0–100)		Bipolar scale (−50 to +50)	
	n	Mean	*n*	Mean
Transparencies and other course material	9	83	12	88
Clarity of instruction	11	82	10	97
Interestingness of subject matter	9	82	10	90
Competence of instructor	12	78	13	94
Relevance for exam	4	78	5	87
Literature (readability, availability)	9	72	5	91
Cooperation between instructor and students	8	62	10	82
Atmosphere/rooms	5	56	3	34
Instructor's appearance	4	10	3	20

Original bipolar scale values are augmented by 50.

5.2. Results

After the completed sheets had been collected, participants were asked about the presumed purpose of the activity. None of them indicated that they had noticed the two different kinds of scales used. This was confirmed when the results were discussed in class.

The criteria that were listed by three or more of the course participants and their respective mean importance ratings are shown in Table 3. Values for the bipolar scale are original values augmented by 50. It is evident from Table 3 that importance ratings were consistently higher when the bipolar scale was used.

A summary of the distribution of average ratings is shown in Fig. 6. The corresponding effect size that expresses the difference in average ratings between the unipolar and the bipolar conditions is $r = .18$ ($p = .48$). This value is, however, strongly biased by two criteria that received extremely low importance ratings. Without the outliers, the effect is much higher: $r = .75$ ($p = .002$).

5.3. Discussion

This study was done in a course on evaluation research and one topic of the current session was evaluation criteria. Participants had to produce such criteria themselves and therefore one should expect a rather high level of involvement

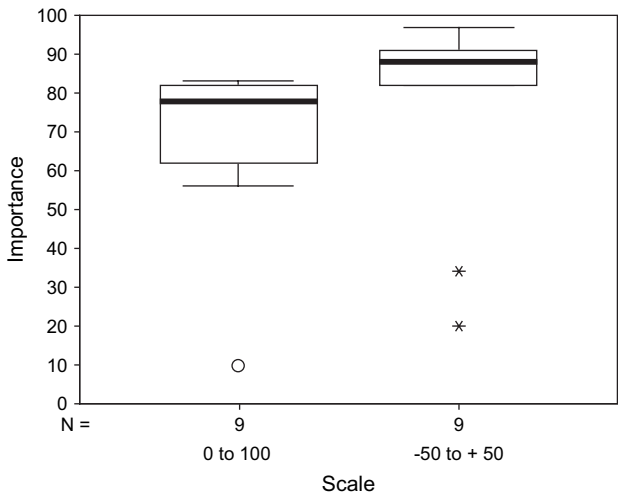

Fig. 6. Box plots showing the distributions of average importance ratings for nine evaluation criteria dependent on endpoint numbering of scales. The left distribution depicts the unipolar and the right the bipolar scale results.

in the task. High involvement in turn might lead to ratings that are less influenced by the endpoint scaling than by the position on the scale. Nonetheless, the effect found in this study was, although not directly comparable to the effects in the prior studies, definitely not less pronounced. High involvement, either in the content of the course or in the rating task itself, does not seem be a sufficient condition to prevent the effects of scale polarity found in all the present studies.

6. General discussion

Four studies examined the role of scales on ratings of several aspects of course evaluations. Consistent with prior survey research (Schwarz, 1999) and research on the construction and use of tests (Attali & Bar-Hillel, 2003) these studies confirmed that the way a scale is constructed may also have a strong impact on course-related student ratings. The effects reported here focused on three possible variations of scales: endpoint numbering (unipolar vs. bipolar scales), different ranges within scales (variation in the length of time intervals offered) and the ordering of choices.

Bipolar scales generally elicited higher values on the transformed scale than unipolar scales. It seems that if participants have a positive attitude towards an issue, they tend to use only the positive part of the bipolar scale. If the bipolar values are then transformed (e.g., by adding 50 to the values elicited by a −50 to +50 scale), average ratings can be expected to be markedly higher than when a unipolar scale is used. This effect seems to be influenced by how certain one is about the issue but apparently not by how much one is involved in it or in the rating task itself.[2] This finding does not preclude that long-term involvement in an issue might eventually lead to higher certainty and therefore reduce scale dependent variations. Moreover, it remains to be examined whether the trend found works in both directions. If the current interpretation of the effect holds, participants' prevailing attitude toward the issues asked was generally positive and therefore, they tended to use only the positive part of the bipolar scale. If the general attitude was negative, there should be a tendency to use only the negative part of a bipolar scale.

The variation of time intervals within a scale that consists of rank ordered options indicates that participants tend to orient themselves on the scale values as a frame of reference (Schwarz et al., 1985). Participants in the current studies might have taken the time intervals in the middle of the scale as the usual amount of time spent for a session of the course. One might speculate whether this effect breaks down when these middle values become unrealistically large (e.g. "up to 10 h," "10–20 h,"…).

The most subtle manipulation of scales in the above studies consisted in varying the order of alternatives. The effects obtained with this variation were not very large but consistent over two studies: options displayed in the middle positions of a list of evaluation criteria were preferred over options displayed on the outside positions of that list. Although small, the effect would have sufficed to change the rank order of the judged importance of some evaluation criteria.

6.1. Relevance of results

The effects found in the present studies, at least those addressing the impact of endpoint labeling and of different ranges within scales, are rather large as compared to effects usually found in psychological research (Cohen, 1992). But apart from their size, are the findings also relevant? I think so, because they represent some important aspects of course evaluation reports and may have a practical impact on the comparison and interpretation of course evaluations and course-related judgments.

The studies reported did not deal with complete course evaluations and, in contrast to most course evaluations, the questions were not asked at the end of the course, thus partly yielding judgments about future behavior. Nonetheless, the questions examined in the present studies address important aspects of the multiple dimensions that determine the complex activity of teaching and that should be reflected in students' course evaluations (Marsh, 1984; Marsh & Roche, 1997). The present results indicate that ratings for the dimensions examined (and probably other ones) can

[2] One might argue that the smaller effects for the interest ratings as compared to the relevance ratings might not have been due to differences in certainty but due to differences in self-relevance (participants need to score highly in their interest ratings because if they were not interested in their (main) subject, what could be the reason for studying it?). Whereas this explanation cannot be fully ruled out with the current experimental design, there is no a priori reason to expect a different impact of self-worth reasons on interest ratings for psychology students (Study 2b) and students who studied for a minor in psychology (Studies 1 and 2a), so their ratings should be comparable, which they are not (compare Figs. 1 and 2 to Fig. 4).

be moved upwards or downwards by choosing a certain type of scale. Moreover, the ratings themselves may be used for subsequent judgments about related issues (Schwarz, 1990, 1999). So if, for instance, a student infers from her relative position on the time scale, that the amount of time she plans to spend on course preparation is well below the ''average,'' as indicated by the range of possible answers, she might conclude that the course is not so important for her, after all, and behave accordingly.

Although the effect of the position of rating alternatives – i.e., whether they appear in the middle or in the outside positions of a list of alternatives – seems to be not very large, it might be the most difficult to avoid. The problem of ordering response alternatives arises in all multiple-choice questionnaires and small effects might add up to substantial distortions even in standardized instruments.

When the results of course evaluations depend on the scale used, this might also have implications for the comparison of evaluation reports across institutions. If, for instance, course evaluations are used when hiring a new faculty member, applicants whose teaching performance is measured by using unipolar scales might fare less well than ones whose performance is measured by bipolar scales when ratings are standardized (e.g. as proportion of maximally obtainable points).[3] Moreover, one might be tempted to argue with ''above average'' or ''below average'' results, taking the middle of the scale as ''average.'' Especially if students are more or less undecided about an issue, results could be easily ''moved'' in one direction or the other. Finally, scale effects might also bias meta-analytic reviews of course evaluation reports and should be regarded as potential moderating variables in such analyses.

6.2. Recommendations

What can be done about the potential influence of scales on the results of evaluation reports and course-related ratings? One possibility would be to leave everything as is but be aware of possible biases and adjust for these biases when making judgments or decisions. This might sometimes be the only way but could turn out to be a difficult task. If questions concern frequency estimates it might be best to ask participants in an open response format. Estimates of frequency and relative frequency have been found to be remarkably accurate (Sedlmeier & Betsch, 2002). Directly asking about an estimate might also be applicable to questions about time. It also helps to make questions more precise (Schwarz, 1999). Finally, and most importantly, one should think about working towards creating standardized modules that are used by all or at least by most universities within a given country.

The bias arising from the ordering of answer alternatives can be reduced by a randomizing procedure when there are many such questions to be asked on one questionnaire (Attali & Bar-Hillel, 2003; Bar-Hillel & Attali, 2002). When a questionnaire only includes one or a few questions in this format, one might just balance the positioning of answer alternatives as in the above studies and then possibly take the (weighted) average if the inspection of results for the middle vs. outside positions does not show anomalies. However, one should always be aware about the possible influence of answer positioning.

As already mentioned, the present research did not deal with full-fledged student evaluations of teaching. Nonetheless, some recommendations for the development of respective questionnaires seem possible in the light of the results. In theory, there is an easy solution to most of the biases introduced by scaling effects: use only one standardized procedure. In practice, a common agreement about which questionnaire is the best one might be difficult to achieve because several groups of researchers might come up with different solutions. The task of finding an (near) optimal procedure might be made easier if a nation-wide institution could bundle all the efforts and could make possible a low cost supply of state-of-the-art questionnaires (which could be periodically updated in response to new insights) to all the respective universities and schools. If one would succeed in that endeavor, most effects found in the present studies would of course not be relevant any longer. However, the problem posed by ordering effects cannot be solved by standardization alone. A possible recommendation for course evaluation forms containing multiple-choice answers would be to offer always two or more parallel forms that differ in their ordering of response alternatives, to apply all the forms for a given evaluation project in equal proportions, and to take the averages across the different forms as the end result.

Students cannot be expected to be really certain about most aspects asked about in course-related questions. The results of the current studies indicate that in this situation they may also rely on the information provided by the scale

[3] There is some potential for that to happen. For instance, two of the three most widely used standardised questionnaires for course evaluations in Germany, the ones of Diehl and Kohr (1977) and Staufenbiel (2000) are quite similar but differ in their scales. Whereas Diehl and Kohr (1977) use a bipolar scale, Staufenbiel employs a unipolar scale that uses comparable labels.

and thereby adjust their ratings accordingly. Course evaluations are necessary and important for the improvement of teaching. Therefore, every effort should be taken to create valid measurement devices. In this effort, special care should be paid to the design of scales.

Acknowledgements

I thank Cornelia Belger, Jane Endler, and Maya Reimer for their assistance in data collection and Tilmann Betsch, Johannes Hönekopp, Frank Renkewitz, Frank Ritter as well as Anita Todd for their feedback on an earlier version of the paper, and I am grateful for the insightful comments of three anonymous reviewers that helped to improve the paper considerably.

References

Abrami, P. C., & d'Appolonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall, & J. Franklin (Eds.), *Student ratings of instruction: Issues of improving practice* (pp. 97–111). San Francisco: Jossey-Bass.

d'Appolonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*, 1198–1208.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: the position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*, 109–128.

Bar-Hillel, M., & Attali, Y. (2002). Seek whence: answer sequences and their consequences in key-balanced multiple-choice tests. *The American Statistician*, *56*, 299–303.

Basow, S. A. (1995). Student evaluations of college professors: when gender matters. *Journal of Educational Psychology*, *87*, 656–665.

Boehnke, K., Petri, K., Baier, D., & Goldschmidt, R. (2000). *Evaluationsprojekt ''Universitätsweite beleglesergestützte Lehrevaluation'' – Ergebnisse der zweiten Befragungswelle*. Available from http://www.tu-chemnitz.de/phil/soziologie/boehnke/.

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology*, *84*, 563–572.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Diehl, J. M. (2001). Studentische Lehrevaluation in den Sozialwissenschaften: Fragebögen, Normen, Probleme. In E. Keiner (Hrsg.) (Ed.), *Evaluation (in) der Erziehungswissenschaft*. Weinheim, Germany: Deutscher Studien Verlag.

Diehl, J. M. (2003). Normierung zweier Fragebögen zur studentischen Beurteilung von Vorlesungen und Seminaren. *Psychologie in Erziehung und Unterricht*, *50*, 27–42.

Diehl, J. M., & Kohr, H. U. (1977). Entwicklung eines Fragebogens zur Beurteilung von Hochschulveranstaltungen im Fach Psychologie. *Psychologie in Erziehung und Unterricht*, *24*, 61–75.

Grice, H. P. (1975). Logic and conversation. In D. Davidson, & G. Harman (Eds.), *The logic of grammar* (pp. 64–75). Encino, CA: Dickenson.

Greenwald, A. G., & Gillmore, G. M. (1997a). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology*, *89*, 743–751.

Greenwald, A. G., & Gillmore, G. M. (1997b). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*, 1209–1217.

Harris, R. J., & Monaco, G. E. (1978). Psychology of pragmatic implication: information processing between the lines. *Journal of Experimental Psychology: General*, *107*, 1–27.

Lebherz, C., Mohr, C., Henning, M., & Sedlmeier, P. (2005). Wie brauchbar sind Hochschul-Rankings? Eine empirische Analyse. *Zeitschrift für Pädagogik. Beiheft*, *50*, 188–208.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*, 560–572.

Marsh, H. W. (1982). SEEQ: a reliable, valid, and useful instrument for collecting student's evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77–95.

Marsh, H. W. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707–754.

Marsh, H. W. (1995). Still weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology*, *87*, 666–679.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. The critical issues of validity, bias, and utility. *American Psychologist*, *52*, 1187–1197.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92*, 202–228.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology*, *82*, 250–261.

Rindermann, H. (2003). Lehrevaluation an Hochschulen: Schlussfolgerungen aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation. *Zeitschrift für Evaluation*, *3*, 233–256.

Rindermann, H., & Amelang, M. (1994). *Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation (HILVE). Handanweisung*. Heidelberg, Germany: Asanger.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russel Sage Foundation.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research*. New York: Cambridge University Press.

Schwarz, N. (1990). Assessing frequency reports of mundane behaviours: Contributions of cognitive psychology to questionnaire construction. In C. Hendrick, & M. S. Clark (Eds.), *Research methods in personality and social psychology. Review of personality and social psychology*, *Vol. 11* (pp. 98–119). Beverly Hills, CA: Sage.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist*, *54*, 93–105.

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response categories: effects on behavioral reports and comparative judgments. *Public Opinion Quarterly*, *49*, 388–395.

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570–582.

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research – Online,* 1. http://www.mpr-online.de/.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah: Erlbaum.

Sedlmeier, P., & Betsch, T. (Eds.). (2002). *Etc. Frequency processing and cognition*. Oxford: Oxford University Press.

Spiel, C., & Gössler, P. M. (2000). Zum Einfluß von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende. *Zeitschrift für Pädagogische Psychologie*, *14*, 38–47.

Staufenbiel, R. (2000). Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. *Diagnostica*, *46*, 169–181.