Linek, Stephanie B.

Evaluation of Civic Science Education: Influence of the Time Point of Measurement

Proceedings of the 16th International Conference on Education and New Learning Technologies (EDULEARN 2024)

Suggested Citation: Linek, Stephanie B. (2024) : Evaluation of Civic Science Education: Influence of the Time Point of Measurement, Proceedings of the 16th International Conference on Education and New Learning Technologies (EDULEARN 2024), ISBN 978-84-09-62938-1, IATED Academy, Valencia, pp. 1771-1779, https://doi.org/10.21125/edulearn.2024.0533

This Version is available at:
http://hdl.handle.net/11108/625

EVALUATION OF CIVIC SCIENCE EDUCATION: INFLUENCE OF THE TIME POINT OF MEASUREMENT

S.B. Linek

ZBW Leibniz Information Centre for Economics (GERMANY)

Abstract

Civic science education is a good way to engage the youth to make new experiences. Thereby, the evaluation of such educational initiatives is important to ensure their effectiveness and sustainability but often has to face the challenge of practical feasibility. An easy way of evaluation is the use of an online survey immediately at the end of an educational initiative. This seems to be a good time point for capturing a snapshot of the children's enthusiasm, and has the practical advantage that the children's willingness to fill in a survey on their experiences is probably relatively high. However, on the other hand, this time point might be inappropriate because the short-term excitement of the children might bias the validity of the results on the long-term benefits of the educational initiative.

Beyond this background, the presented study investigated the influence of the time point of evaluation measurements. Thereby a school competition served as practical example of civic science education. Half of the children received an email-invitation to fill in an online evaluation survey immediately at the end of the school competition (immediately-group), and the other half of the children received the same email-invitation with a delay of four weeks (delay-group). The evaluation survey included a global judgement as well as detailed ratings of single elements of the school competition and of individual values of participation.

The results of the 207 interviews showed that the response rate of the immediately-group was much higher than the response rate of the delay-group. Both groups gave an equally high global rating of the school competition. However, some important single elements and individual values received significantly lower ratings by the delay-group. Overall, the results indicated that the excitement of the children at the end of the school competition and social desirability caused partly a positive bias in the immediately-group.

The findings illustrate the trade-off between practicability and validity. Measurements immediately at the end of an educational initiative have a better practical feasibility and provide a solid global judgement. However, delayed measurements deliver a more valid and less emotional picture of the specific elements and individual values of participation. Thus, a delayed evaluation seems to be more appropriate as a basis for future improvements and as the measurement of the long-term benefits of civic science education.

Keywords: Evaluation, methodology, time point of measurement, civic science education, informal learning, science popularization.

1 INTRODUCTION

Civic science education is an important cornerstone of science communication outside the formal structures of the school system and other official institutions. Civic science education can be described as participative science communication that supports people's ability to understand, explore and take informed action in science-related public issues [1]. Thus, civic science education aims at broadening the people's horizon, deepening scientific understanding, and creating new interactive learning experiences. Accordingly, the concrete educational initiatives of civic science education are usually highly interactive and include a close dialogue and participative exchange of between laypeople and scientific experts.

The highly interactive and participative character of civic science education makes the corresponding educational initiatives very demanding for the participants and the organizers. Correspondingly, also the evaluation of civic science education has some special aspects. Namely, the evaluation of the single elements of a concrete initiative includes also the evaluation of the close cooperation and interaction between laypeople and scientific experts. Further, the sustainability and long-term effects of the scientific experiences relate less to single scientific topics but rather to the people's general ability to understand and explore science, and to take informed action in science-related public issues. Thus, the evaluation

of civic science education is partly a little bit special in relation to the evaluation aims. Nevertheless, most of the methodological issues discussed below can be applied analogously to the evaluation of other forms of science communication and informal learning. In the following, some basic aspects of the evaluation of science communication in general will be outlined. Afterwards, the practical challenges of the evaluation of civic science education and the corresponding educational initiatives will be discussed in more detail.

The evaluation of science communication in general – including civic science education and informal learning – is important to ensure the effectiveness and quality of concrete educational initiatives. Thus, evaluation measurements should be science-based and methodologically reasonable to provide valid evaluation results [2] [3] [4]. This is even more important when the evaluation results serve as the basis for future improvements or should provide general recommendations for the practice of science communication. A well-designed and scientifically based evaluation can require a high amount of time and resources [4] [5]. On the other hand, the evaluation of educational initiatives outside the formal structures of official institutions like schools or universities usually has limited resources. Also, evaluation measurements have to face several challenges of practicability like regulation of privacy, and last but not least, the compliance of the participants. Thus, a reasonable evaluation of science communication have to regard for the balance between scientific accuracy and practical feasibility.

There are several possibilities of measurement, and each method has its advantages and drawbacks [6]. Thereby, the decision on the evaluation method should consider not only the practicability but also the concrete aims of the evaluation. Further, the validity of the evaluation results is decisional to derive accurate implications [7]. To ensure valid results, the evaluator has also to take care of potential method biases, i.e., systematic error variance that traces back to the applied measurement method and might bias the results. Method biases can have a severe impact on the validity of the results and thus, could lead to misleading conclusions. Podsakoff and colleagues [8] provide a good systematic and very detailed overview on possible sources of method biases. For example, the items of a questionnaire, their order, and their wording are possible sources of biases. Also, the answering style of the respondents (e.g., acquiescence bias) should be controlled and considered. In this relation, a careful design of the measurement instrument is a critical factor, and often it is advantageous to use standardized or already tested scales and questionnaires [8] [9] [10].

Besides an appropriate measurement instrument, there are also biases due to the context, circumstances, and individual characteristics of the participants. Some of them are hard to control when evaluating science communication. In the following, I describe exemplarily some important biases that are of special interest for the evaluation of civic science education because of the participative and highly interactive nature of the corresponding educational initiatives.

A prominent example of a method bias is social desirability. Social desirability can be generally described as the individual tendency to present oneself in a favourable direction [11] [12] [13] [14]. In the context of civic science education, the social desirable behaviour is to present oneself as interested in science and highly motivated. Further, the social norm is to be a team player, open-minded, and to behave friendly and cooperative towards scientific experts, researchers, and educators. Additionally, some people want to be a “good participant” who gives the answers, the experimenter expects. Such behaviour could be even more pronounced in face-to-face interviews or if the experimenter is present [15]. This is quite close to the courtesy bias, i.e., the tendency to give polite answers instead of honest answers that seem to be rude [16]. A similar bias can occur in the form of a leniency bias if people have to give a rating about another person. The leniency bias denotes the tendency to assign more positive ratings to a person if he/she is liked (compared to a disliked person) [17] [18]. The courtesy bias and the leniency bias illustrate that biases are partly interconnected and it can be hard to separate them from each other. As pointed out above, it is the social norm in civic science education to be a convenient person with an open mind towards science and to behave friendly towards the scientists who dedicated their time to the discourse with laypeople. Thus, in the context of civic science education, the leniency bias and the courtesy bias can be seen as a sub-form of social desirability.

Besides social desirability, also the mood state of the respondents can bias their answers. Thereby not only the general negative of positive affective disposition can influence the answers of a person, but also very transient mood states like short-term excitement or just having a bad day [8]. Additionally, contextual influences like the time, location, and used media of measurement can result in systematic method biases. Thereby, contextual factors can also provide contextual cues that in turn elicit other biases like the mentioned bias by the transient mood state of the participants. Similarly, the demanded characteristics of the setting can influence the answers. For example, the presence (versus absence) of the person that should be evaluated (e.g., teacher that should be evaluated by his/her students) could

induce the social pressure to give polite evaluations and thus can lead to a social desirability bias (in form of courtesy or leniency).

Accordingly, also the time point of measurement is an important decision. For example, (delayed) retrospective surveys bear the danger of retrospective memory biases [19] [20] [21] [22]. However, also non-retrospective surveys have their pitfalls, and retrospective surveys are often undervalued [6]. Additionally, non-retrospective measurements are partly not practicable, e.g., surveys on health care services can hardly be applied during the emergency care of an ill person. Similarly, in marketing research, consumer behaviour can often only be measured with some delay [6]. Further, non-retrospective measurements are often connected with the presence of the interviewer and this in turn can influence the answers of the participants (interviewer bias). Finally, yet importantly, immediate non-retrospective measurements are often not appropriate when the evaluation relates to long-term effects, because instant questions can only assess answers on the assumed sustainable effects. Similarly, measuring the learning success immediately after a lesson delivers different results compared to a delayed measurement of the long-term learning benefits after some months.

There exist several possibilities to control such method biases [8] [12] [23], but they are often very resource-intensive, need additional time for preparation, and require highly compliant participants. Thus, often a compromise is needed between the practical feasibility and scientific methods that assure validity.

Online surveys with multiple-choice questions and rating scales can be a good starting point. The questions can be carefully prepared in advance and thus, method biases by the wording and the arrangements of the items can be avoided or controlled. In the best case, an already tested questionnaire is available. Additionally, the data of online surveys can be comparable and easily analysed (without additional transcription or coding) and potential response biases of the participants become often obvious during the data analysis (e.g., acquiescence bias).

Nevertheless, the critical point is to ensure the compliance of the participants to fill in the survey. Thereby, the context and the time point of evaluation measurements are important aspects. For many informal educational initiatives, it is not feasible to have pre- and post-measurements (due to limited resources and limited compliance of the participants). Partly, the only chance to collect evaluation data is immediately at the end of the initiative because later on it is not possible to interrogate the participants due to lacking contact information.

Thus, at first glance, it seems reasonable to collect the data immediately at the end of an initiative when children are still present and available. At this time point, the children can be easily motivated to answer questions on the educational experiences they just had. In principle, immediately at the end of an initiative is a good time point to capture the spirit of an educational initiative. Also, the participants can easily remember their educational and scientific experiences, that mean, the instant answers are probably not affected by retrospective memory biases.

On the other hand, immediately at the end of an initiative, it is simply not possible to measure the long-term value and the sustainability. At this time point, the participants do not know how much they will forget and if and how long and to which extent they will profit from their educational and scientific experiences later on. In comparison, after some delay, the participants already might have recognized how fast they will forget scientific details or how much they still profit from their participation after everything was finished. Thus, after some delay, they could probably better estimate how long the benefit of the educational initiative will last.

Additionally, immediately at the end of an initiative there are usually some situational factors that might affect the validity. This is especially true for civic science education with its highly interactive nature. First, measurements made immediately at the end could be emotionally biased by the short-term excitement (enthusiasm or frustration when the scientific experience and the interaction with scientists is just finished) of the participants. That means, there could be a bias by the transient mood state of the participants, and thus could differ substantially from judgements made with some delay. This is even worse when the evaluation aims relate to the long-term effects and sustainability of civic science education.

Also the influence of social desirability could be stronger immediately at the end of an educational initiative if the organization team and the involved scientists (which are usually also part of the evaluation) are still present. In civic science education, it is highly socially desirable to stand behind the own educational or scientific participation and to be enthusiastic. The presence of important key persons (scientists, educators, and organizers) makes this social norm more salient; thus, people might act more politely and in a socially desirable way. Similar, if the participants are (immediately at the end) still in close interaction with their peers, they are in a social context with highly salient social norms, for example

being friendly, being a team player, and show enthusiasm about the made experiences. Thus, a delayed measurement could be more appropriate to deliver a down-to-earth judgement and less polite answers.

To sum up, the time point of evaluation measurements might be a critical factor of influence that is connected with several potential method biases. Thus, one has to assure that the chosen time point of measurements delivers valid evaluation results. Thereby, also the aims of evaluation should be considered. Is it just a global evaluation in the sense of a holistic judgement? Or is the evaluation conceptualized as the basis for specific future improvements and to receive a detailed picture of the long-term sustainability of civic science education?

Based on these considerations, the following study investigated the influence of the time point of evaluation measurements. More specifically, the research question was, if and how the time point of the presentation of an online survey influences the evaluation results of an educational initiative of civic science education.

2 METHODOLOGY

2.1 Setting of the Study

To address the research aim, the study used a school competition as a case of an educational initiative of civic science education, namely the annual YES! (Young Economic Solutions) school competition in 2022. The YES! school competition on economic, social, and environmental challenges of the future (https://young-economic-solutions.org/en/) followed a participative approach and was based on several elements that support the children's ability to understand science and develop their own solutions to science-related challenges of the future by the help of scientists. (Further details on the YES! project in general was described by Linek and Scholz [24].) At the end of the school competition, the participants were invited to fill in an evaluation survey for a continuous monitoring of the effectiveness of the YES! project and as the basis for future improvements. Most questions of the evaluation survey were already tested in former evaluation studies of the YES! project. It comprised not only ratings on the school competition and its elements as basis for future improvements, but also included questions on the personal values of participation to receive additional insights in the motives of the children and in the general benefits of civic science education (see [25]). The evaluation survey was organized as online questionnaire and the participants received the invitation to fill in the survey by email. The email-invitation was sent out by the same mailing list that served as the internal communication during the school competition for providing relevant information.

2.2 Design of the Study

To answer the research question on the influence of the time point of the evaluation measurements, a two-group design was used. Both groups received the same online survey with identical questions, but the time point of measurement was different. Half of the children received an email-invitation to fill in the online survey immediately at the end of the school competition (immediately-group), and the other half of the children received the analogous email-invitation about four weeks later (delay-group). The two groups were balanced as far as possible with respect to school type, county of the schools, size of the school teams, and age and gender of the children. The participation was voluntary. As a reward, the children could take part in a lottery.

2.3 Variables

Variables were a global judgement as well as detailed ratings of single elements of the school competition. The complete list of the included single elements is depicted in Table 1 in the results section. Further, also individual values of participation were investigated as indicators of sustainability in relation to the global aims of civic science education, i.e., the support of the participants' general ability to understand and explore scientific topics, and to take informed action in science-related public issues in the future. These individual values comprised very different aspects ranging from "benefit for my later professional career" to "value for my possibilities to actively participate in shaping the future myself and to bring about a change". Table 2 in the results section shows the complete list of all measured individual values.

2.4 Measurement Instrument: Description of the Online Survey

The online survey was based on prior versions of the regular online evaluation survey for a continuous monitoring of the annual school competition. Thus, the wording of most questions and items was already tested. Only slight modifications of the wording and the items were made in relation to minor practical updates of the annual school competition and correspondence of the research aim of this study. The online survey comprised mainly multiple-choice questions and rating scales. Partly, there were also some open questions and the opportunity to leave open comments to receive additional qualitative insights. At the beginning of the survey, the basic demographics were measured (age, gender, school type, etc.). Afterwards, the questions on the evaluation of the school competition followed. Some of these questions related to very detailed internal practical issues (e.g., if the children have met for team work in private rooms or their school building) and will not presented here. The described variables in form of the global judgements, the ratings of the single elements, and the ratings of individual values were the core questions of evaluation. These variables were measured by the help of 7-point Likert rating scales ranging from 1 to 7. Low ratings reflected a negative evaluation; high ratings reflected a positive evaluation.

3 RESULTS

3.1 Description of the Sample and Response Rate

Altogether 344 children (from German schools) participated in the school competition. Most of the participants were from high school and were between 16-17 years old. 167 children were invited to fill in the evaluation survey immediately at the end of the school competition, and 177 participants were invited with a delay of 3-4 weeks.

Overall, 207 children (106 female, 99 male, 2 no answers) filled in the survey. The children were free to omit single questions, and to skip the survey at each point if they wanted. Thus, the number of valid cases was partly lower for some questions. Out of the 207 interviews, 119 interviews were from the group interrogated immediately at the end of the school competition (immediately-group), and 82 interviews were from the group interrogated with a delay of 3-4 weeks (delay-group). That means, the response rate of the immediately-group was substantially higher (72%) than the response rate of the delay-group (46%).

3.2 Results on the Influence of the Time Point of Measurement

The influence of the time point of measurement was analysed by a statistical comparison of the two groups (immediately-group versus delay-group) by analyses of variance. For the global judgement of the school competition, a one-way analysis of variance (ANOVA) was calculated. For the questions with multiple items, i.e., the ratings of the single elements and individual values, a multivariate analysis of variance (MANOVA) was calculated to account for the interdependences (significant correlations) between the single items.

The overall *global rating* of the school competition on the 7-point Likert scale was relatively high (M = 5.38, SD = 1.25; n = 200). There were no significant differences between the immediately-group (M = 5.45; SD = 1.24; n = 118) and the delay-group (M = 5.28; SD = 1.25; n = 82).

The *single elements* of the school competition received medium to good evaluations. The descriptive statistics of the ratings (on the 7-point Likert scale) of the single elements are listed in table 1. Most of the single elements received comparable high ratings in the immediately-group and the delay-group. However, the results of the MANOVA showed that some of the single elements of the school competition received significantly lower ratings by the delay-group. Namely, teamwork in their own group (F = 13.475; p < .001; *partial Eta²* = 0.121), kick-off with scientists (F = 7.656; p = .007; *partial Eta²* = .072), and that pupils can develop their own solutions (with the help of researchers) on a scientific basis (F = 7.215; p = .008; *partial Eta²* = .069) were rated lower by delay-group compared to the immediately-group. Additionally, there were non-significant tendencies for lower ratings of the expert talk with scientists (F = 3.680; p = .058; *partial Eta²* = .036) and that pupils take the lead in working on their own ideas (F = 3.059; p = .083; *partial Eta²* = .030).

Table 1. Means and standard deviations (in brackets) for the single elements, structure by the time point of measurement (groups). An asterisk marks items with a significant difference between the groups.*

Elements of the school competition	***Immediately (n = 65)***	***Delay (n = 35)***	***All (n = 100)***
Teamwork in the own school team *	5.83 (1.14)	4.70 (1.82)	5.45 (1.50)
Web-seminar about the competition	4.91 (1.48)	4.50 (1.42)	4.78 (1.46)
Kick-off with scientists *	5.28 (1.43)	4.40 (1.52)	4.98 (1.51)
Expert talk with scientists	5.17 (1.66)	4.50 (1.56)	4.94 (1.65)
Regional final	5.78 (1.13)	5.50 (1.22)	5.68 (1.16)
General support for students by researchers	4.83 (1.52)	4.40 (1.63)	4.67 (1.56)
Supporting pupils by the learning modules	4.46 (1.86)	4.20 (1.49)	4.37 (1.74)
Personal guidance and support of pupils by teachers	5.02 (1.61)	4.90 (1.91)	4.96 (1.71)
Coordination by the organization team	4.95 (1.54)	4.90 (1.49)	4.95 (1.51)
Contact to the organization team	4.92 (1.65)	5.10 (1.33)	4.97 (1.54)
Personal guidance and support by the organization team	4.80 (1.62)	4.90 (1.34)	4.84 (1.52)
Pupils take the lead in working on their own ideas	5.80 (1.33)	5.30 (1.32)	5.63 (1.34)
Pupils can develop their own solutions (with the help of researchers) on a scientific basis *	5.82 (1.18)	5.10 (1.61)	5.55 (1.39)
Pupils have the opportunity to exchange with scientists on eye level	5.55 (1.53)	5.30 (1.62)	5.47 (1.56)
Pupils have the opportunity to consult researchers repeatedly	5.18 (1.64)	5.10 (1.83)	5.17 (1.70)
Competition between the school teams	5.20 (1.39)	4.90 (1.71)	5.09 (1.51)
Presentation of the own developed ideas and approaches during the final	5.49 (1.24)	5.30 (1.45)	5.44 (1.31)
Presentation in English (experience with foreign language)	4.49 (1.96)	4.70 (1.97)	4.55 (1.96)
Exchange between the school teams at the final	4.78 (1.60)	5.30 (1.50)	4.95 (1.57)
Presentation in front of researchers and politicians.	5.58 (1.52)	5.40 (1.52)	5.51 (1.51)
Incentive through the prize for the winning team	5.35 (1.61)	4.90 (1.50)	5.18 (1.58)
Incentive through the public honouring of the winning team	5.62 (1.68)	5.50 (1.56)	5.58 (1.63)
Incentive to receive a certificate of attendance for CVs	5.74 (1.58)	5.60 (1.52)	5.68 (1.56)
Free participation (incl. travel and accommodation costs)	5.78 (1.65)	5.40 (1.79)	5.66 (1.70)

The analysis of the *individual values* of the school competition showed a similar pattern like the ratings of the single elements. All individual values were rated on a medium to good level. Thereby, most of the values were rated on a comparable high level in both groups, but some of the individual values received significantly lower ratings if the evaluation measurements were made with delay. The descriptive statistics for the ratings (on the 7-point Likert scales) of the individual values are listed in table 2.

Table 2. Means and standard deviations (in brackets) for the individual values, structure by the time point of measurement (groups). An asterisk marks items with a significant difference between the groups.*

Individual values of the school competition	***Immediately (n = 90)***	***Delay (n = 55)***	***All (n = 145)***
Value for my personal development	5.20 (1.38)	5.27 (1.57)	5.23 (1.45)
Benefit for my later professional career	4.66 (1.57)	4.53 (1.74)	4.61 (1.63)
Broadens my horizon	5.69 (1.31)	5.82 (1.04)	5.74 (1.21)
Value for my idealistic goals	4.74 (1.64)	4.64 (1.61)	4.70 (1.63)
Benefits for society *	5.51 (1.40)	4.64 (1.80)	5.18 (1.61)
Benefits for scientific progress	4.79 (1.60)	4.56 (1.74)	4.70 (1.65)
Benefits for the transfer of knowledge between school and science *	5.38 (1.54)	4.71 (1.80)	5.12 (1.67)
Benefits for economic education outside the classroom	5.26 (1.50)	5.24 (1.44)	5.25 (1.47)
Value for my ability to work in teams	5.67 (1.37)	5.49 (1.36)	5.60 (1.36)

Value for my ability to communicate with experts and scientists	5.40 (1.48)	4.93 (1.61)	5.22 (1.54)
Boost for my general self confidence	5.22 (1.56)	4.95 (1.77)	5.12 (1.64)
Boost for my ability to understand scientific contexts	5.08 (1.39)	4.93 (1.60)	5.02 (1.47)
Value for my possibilities to actively participate in shaping the future myself and to bring about a change	5.28 (1.38)	5.24 (1.58)	5.26 (1.45)
Benefit to deal with relevant topics	5.70 (1.30)	5.53 (1.32)	5.63 (1.31)
Value of competing with others *	5.12 (1.46)	4.60 (1.62)	4.92 (1.54)
Value of communicating and exchange with others	5.27 (1.42)	5.45 (1.40)	5.34 (1.41)

The results of the MANOVA showed that the benefits for society (F = 10.697; p = .001; *partial Eta²* = .070), benefits for the transfer of knowledge between school and science (F = 5.648; p = .019; *partial Eta²* = 0.038), and the value of competing with others (F = 4.024; p = .047; *partial Eta²* = 0.027) were rated significantly lower by the delay-group compared to the immediately-group. In addition, there was a non-significant tendency for lower ratings of the delay-group for the value of the ability to communicate with experts and scientists (F = 3.253; p = .073; *partial Eta²* = .022).

4 CONCLUSIONS

To sum up the findings and to answer the research question of the study, the time point of measurement did not affect the rough global evaluation result, but had partly significant influence on the detailed evaluation of single elements and individual values for participation.

That means, on the one hand, the overall holistic evaluation was rather independent of the time point of measurement. The global judgement and most of the ratings of the single elements and individual values were not influenced by the time point of measurement. Especially the structural single elements of the school competition (like the interaction with the organization team and the possibility to win a prize) were rated on an equally high level. Also, most of the main individual values (e.g., broadening the horizon, personal development, value for idealistic goals or benefit for the later professional career) were rated on the same level. Thus, at first glance, the higher response rate of the immediately-group suggests that measurements immediately at the end of an educational initiative have a better practical feasibility and nevertheless deliver a solid global evaluation.

However, on the other hand, the detailed evaluation measurements on single elements and individual values were partly influenced by the time point of measurement, and namely, some important judgements of the sustainability of the educational initiative received less positive evaluations by the delay-group. Thereby, an interesting pattern appears in the form that all measurements influenced by the time point, correspond to two main aspects.

One aspect related to teamwork and competition. Namely, the teamwork in the own school team, the own lead when working on the own ideas, and the competition with others were rated lower from a distance of time. This could be interpreted in the sense that the excitement and euphoric emotions of the children (due to the close interaction with their peers and scientists) cooled down, i.e., the transient mood bias disappeared. An alternative interpretation could be social desirability. Immediately at the end of the school competition, the teammates and the other competing teams were still present, and this might have produced social pressure to be nice and polite. In contrast, during the delayed evaluation measurement there was less (or no) social pressure by the presence of the other pupils, and thus it was much easier to give a less polite but more honest and rational answer.

A second aspect was the co-working with scientists, and the exchange between science, schools and society. In particular, the delay-group gave lower ratings for the kick-off with scientists, and that people can develop their own solutions with the help of researchers on a scientific base. Further, the benefits for society and the benefits of the transfer between school and science were rated lower by the delay-group. These findings can be interpreted analogous to the findings on teamwork. Immediately at the end of the competition the “spirit” of the competition and the excitement of pupils was rather high and thus, the ratings might be overly euphoric (transient mood bias). In the delay-group, the emotions had already cooled down and thus, the ratings are more down-to-earth. Similar, social desirability could have been a factor of influence. During the immediate evaluation measurement, the social pressure by the presence of scientists might have positively biased the ratings on the support by scientists and the judgements on the assumed exchange between schools and society as well as the benefit for society. An alternative (or additional) explanation could be that after some time delay, pupils recognized their

knowledge gaps when remembering the own work and the learned scientific insights, and thus, adjusted their estimation of the individual values to a more down-to-earth level. Besides, it is worth noticing that the participation in the evaluation survey was voluntary, and thus, it might be that very frustrated pupils did not participate in the evaluation survey.

Overall, the findings indicate that evaluation measurements immediately at the end of an educational initiative have partly a positive bias by social desirability and the transient mood state of the participants. In contrast, delayed evaluation measurements reflect judgements that are less polite and more prosaic. Nevertheless, the global judgement was on a highly positive level for both time points of measurement, which indicate that the lower ratings of the delayed evaluation measurements were not a sign of frustration. Rather, single elements and individual values were judged more critical and differentiated in the sense of a rational adjustment. This suggests that a delayed evaluation provides a more down-to-earth and less emotional picture of the long-term benefits of civic science education.

Thus, as practical outlook, delayed evaluation measurements are more appropriate as basis for concrete improvements of educational initiatives, and for the assessment of the long-term benefits and sustainability of civic science education. However, it has also to be said that the response rate was substantially lower for the delayed evaluation. Accordingly, the findings illustrate the trade-off between practical feasibility and validity. Measurements immediately at the end of an educational initiative have a better practical feasibility and provide a solid global judgement. However, if accurate and less emotional judgements are needed for future improvements and for evaluating the sustainability of civic science education, then delayed evaluation measurements are more appropriate.

ACKNOWLEDGEMENTS

I want to thank all participants of the YES! school competition for their patience and their dedicated time to fill in the online survey. Additionally, I thank the YES! organization-team at the ZBW (www.zbw.eu) and at the JHS - Joachim Herz Stiftung (www.joachim-herz-stiftung.de) for their support. A special thank goes to Andrea Schmidt from the YES! organization team at the ZBW for her kind and conscientious assistance.

REFERENCES

[1] B. L. M. Levy, A. W. Oliveira, and C. B. Harris, “The potential of civic science education: theory, research, practice, and uncertainties,” *Science Education*, vol. 105, pp. 1053-1075, 2021.

[2] E. Jensen, “The problems with science communication evaluation,” *Journal of Science Communication*, vol. 13, no. 1, C04, 2014.

[3] E. Jensen, “Highlighting the value of impact evaluation: enhancing informal science learning and public engagement theory and practice,” *Journal of Science Communication*, vol. 14, no. 3, Y05, 2015.

[4] R. Ziegler, I. R. Hedder, and L. Fischer, „Evaluation of Science Communication: Current Practices, Challenges, and Future Implications,” Frontiers in Communication. 6:669744, 2021. Retrieved from https://www.frontiersin.org/articles/10.3389/fcomm.2021.669744/full

[5] S. Spicer, “The nuts and bolts of evaluating science communication activities,” *Seminars in Cell & Developmental Biology*, vol. 70, pp. 17-25, 2017.

[6] R. East and M. D. Uncles, “In praise of retrospective surveys,” *Journal of Marketing Management*, vol. 24, no. 9-10, pp. 929-944, 2008.

[7] S. Wolming and Ch. Wikström, “The concept of validity in theory and practice,” *Assessment in Education: Principles, Policy & Practice*, vol. 17, no. 2, pp. 117-132, 2010.

[8] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, “Common method biases in behavioral research: A critical review of the literature and recommended remedies,” *Journal of Applied Psychology*, vol. 88, no. 5, pp. 879-903, 2003.

[9] D. F. Alwin and J. A. Krosnick, “The reliability of survey attitude measurement: the influence of question and respondent attributes,” *Sociological Methods & Research*, vol. 20, no. 1, pp. 139-181, 1991.

[10] R. A. Peterson, *Constructing effective questionnaires*. Thousand Oaks, CA: Sage, (2000).

[11] R. J. Fisher and J. E. Katz, “Social-desirability bias and the validity of self-reported values,” *Psychology & Marketing*, vol. 17, pp. 105-120, 2000.

[12] A. Furnham, “Response bias, social desirability and dissimulation,” *Personality and Individual Differences*, vol. 7, no. 3, pp. 385-400, (1986).

[13] D. Crowne and D. Marlowe, *The approval motive: Studies in evaluative dependence*. New York: Wiley, 1964.

[14] D. L. Paulhus, “Socially desirable responding: The evolution of a construct,” In *The role of constructs in psychological and educational measurement* (H. I. Braun HI, D. N. Jackson, and D. E. Wiley, eds.), pp. 49-69, Mahwah NJ: Erlbaum, 2002.

[15] W. L. Richman, S. Kiesler, S. Weisband, and F. Drasgow, “A meta-analytic study of social desirability distortion in computer administered questionnaires, traditional questionnaires, and interviews,” *Journal of Applied Psychology*, vol. 84, pp. 754-775, 1999.

[16] P. Glick, “How reliable are surveys of client satisfaction with healthcare services? Evidence from matched facility and household data in Madagascar,” *Social Science & Medicine*, vol. 68, no. 2, pp. 368-379, 2009.

[17] J. P. Guilford, *Psychometric methods (2nd ed.)*. New York: McGraw-Hill, 1954.

[18] C. A. Schriesheim, A. J. Kinicki, and J. F. Schriesheim, „The effect of leniency on leader behavior descriptions,” *Organizational Behavior and Human Performance*, vol. 23, pp. 1-29, 1979.

[19] L. Hipp, M. Bünning, S. Munnes, and A. Sauermann, “Problems and pitfalls of retrospective survey questions in COVID-19 studies,” *Survey Research Methods*, vol. 14, no. 2, pp. 109-114, 2020.

[20] J. Bound, C. Brown, and N. Mathiowetz, “Measurement error in survey data.” *Handbook of Econometrics*, vol. 5, chapter 59, pp. 3705–3843, 2001.

[21] J. Pina Sánchez, J. Koskinen, and I. Plewis, “Measurement error in retrospective work histories,” *Survey Research Methods*, vol 8, no. 1, pp. 43-55, 2014.

[22] E. Jaspers, M. Lubbers, and N. D. D. Graaf, “Measuring once twice: An evaluation of recalling attitudes in survey research,” *European Sociological Review*, vol. 25, no. 3, pp. 287-301, 2009.

[23] A. Nederhof, “Methods of coping with social desirability bias: a review,” *European Journal of Social Psychology*, vol. 15, no. 3, pp. 263-280, 1985.

[24] S. B. Linek and W. Scholz, “Participative science communication and the practical use case of the YES!-project,” *Proceedings of the 12th International Conference on Education and New Learning Technologies (EDULEARN 2020)*, pp. 4612-4621, 2020.

[25] S. B. Linek and A. Schmidt, „Parental background and children’s view of civic science education - narrowing the education gap by idealism?” *Proceedings of the 15th International Conference on Education and New Learning Technologies (EDULEARN 2023)*, pp. 2550-2558, 2023.