Online Resource 1. Applied methodology to measure cognitive learning[1]

Measuring perspectives

Among the methods that can be used for identifying individual perspectives are interviewing (e.g., Denzin and Lincoln 2000), cognitive mapping (e.g., Eden 1988; Ridder et al. 2005), and card sorting (e.g., Pahl-Wostl and Hare 2004; Rugg and McGeorge 1997). Traditionally, these methods were used to elicit expert knowledge, but they can also be used for obtaining lay knowledge (cf. Evans 1988), values and interests. To identify major knowledge gaps and conflicts of interest, and to support presentation and discussion of perspectives, it is useful to summarize differences and similarities by grouping individual perspectives. Compared to other elicitation methods, Q methodology supports an objective and reproducible grouping of perspectives relatively well. A strength of Q methodology is that it does not require shared perspectives, or groups of subjects that share them, to be known or hypothesized in advance (Donner 2001). Moreover, it analyses each individual perspective as a whole, and does not aim to generate correlations between objective attributes that are abstracted from the individual (Steelman and Maguire 1999), such as nationality, gender, age and preferred management strategy. Q sorting can be performed by a small, selected sample of individuals and is not intended to generalize the results to a larger population (Steelman and Maguire 1999).

Q methodology requires careful interpretation of sophisticated statistical results (Rugg and McGeorge 1997). Therefore, a new Q analyst should do some reading in order to be introduced in the methodology. Below, we describe the five basic steps in a Q methodological study (cf. Donner 2001; van Exel and de Graaf 2005), as well as the main choices we made in the application of Q methodology in the cases.

1. Collection of all possible statements about the issue at stake (the "concourse")

We collected the concourse of statements by means of interviewing relevant stakeholders and studying policy documents, newspapers and scientific literature. This way, we developed a broad concourse that contained elements of the perspectives of all stakeholders. In order to improve the understandability and recognizability of the statements, we kept the formulation of the statements close to the original formulation.

2. Selection of most relevant statements (the "Q set")

The selection of the most relevant statements from the concourse is a crucial activity in Q methodology. No matter what effort is undertaken, however, obtaining a balanced set remains "more an art than a science" (Brown 1980). The selection can be done according to a fixed structure - either imposed on the concourse (e.g., Dryzek 1993) or emerging from it - or in a more intuitive way. In the case studies we used an intuitive or bottom-up approach. We are of the opinion that, when a practical management issue is at stake, such an approach is more useful than a strict theoretical framework, because it allows for including the most relevant aspects from a practical management point of view. The number of statements in the Q set usually varies between 40 and 60, depending on the complexity of the issue at stake and on the time the respondents are willing to spend. In the cases, we kept the number of statements close to 40, in order to limit the time required for the sorting. At the same time, we tried to select Q sets that were broad and clear enough to activate the tacit criteria, or underlying values, of all respondents (cf. Donner 2001). We selected in particular statements on which opinions were expected to diverge. We discussed preliminary Q sets with colleagues and with the most strongly involved stakeholders in the cases and adapted the Q sets accordingly. Finally, we edited the statements and inserted them in an online Q sorting tool[2].

3. Selection of respondents (the "P set")

The P set should be a structured sample of relevant stakeholders who may be expected to have clear and distinct viewpoints. The P set should maximize the likelihood that all major perspectives on the issue are included (Brown 1980). The number of respondents is usually between 20 and 40. In the Lower Rhine case,

[1] This Online Resource is Electronic Supplementaty Material to the article "Learning from collaborative research in water management practice" published in the journal Water Resources Management (WARM). The material originates from section 4.3 of the PhD thesis "Does collaboration enhance learning? The challenge of learning from collaborative water management research", which was written by the first author of the article in WARM. The thesis was published by VSSD and can be freely accessed online at http://www.vssd.nl/hlf/f042.htm or http://repository.tudelft.nl/view/ir/uuid%3Ab971f82b-bab1-4978-8f35-aecf2d69bcd2/.

[2] Freely available at http://q.sortserve.com/.

we addressed a large and varied group of organized stakeholders that we found to be involved in the issue at stake. In the Delft case, however, we addressed only the stakeholders that were already involved in the ongoing collaborative research process.

4. Ranking of statements by respondents ("Q sorting")

We addressed the stakeholders in the P set by email and asked them to complete the Q sorting online[3]. The respondents were instructed to rank the statements in the Q set according to their personal agreement with each statement, by assigning a fixed number of statements to seven score categories. This resulted in a fixed, uni-modal, and symmetric distribution of statements over score categories (see Table 4.2). Such a fixed distribution forces respondents to carefully compare the statements relatively to each other. This is assumed to decrease the risk of arbitrary or biased sorting, for example under influence of the respondent's mood at the time of sorting, and thus to increase the repeatability of the sort. However, respondents may be dissatisfied about the time and effort required to iteratively put a fixed number of statements in each score category, and about the fact that their perspective cannot be expressed well using the given distribution (cf. Rugg and McGeorge 1997, who see this as a major disadvantage of Q sorting). Such dissatisfaction could be prevented by allowing respondents to freely distribute statements over score categories, without prescribing the shape of the distribution (e.g., Steelman and Maguire 1999). This has no significant consequences for the factor analysis (McKeown and Thomas 1988). When respondents are not at all stimulated to evaluate their agreement with one statement relatively to their agreement with another, however, accuracy of the elicited perspectives will be low.

Table 4.2. Example of fixed distribution of statements over score categories

Meaning	Most disagree						Most agree
Score category	-3	-2	-1	0	1	2	3
Number of statements	4	5	9	10	9	5	4

We used an online Q sorting tool, because it allowed respondents to perform the sort at any convenient time. Furthermore, it significantly reduced the time that we needed to conduct the sort. Disadvantages of an online set-up are the potentially lower response rate, the limited possibilities to explain respondents how to perform the task and the limited flexibility to deviate from the fixed score distribution. There is no apparent difference, however, in reliability and validity of computer- and interview-based Q sorts (van Tubergen and Olins 1979 in van Exel and de Graaf, 2005) .

Before the actual ex ante Q sorting we asked the respondents some questions about their background. Directly after the ex ante Q sorting, we asked the respondents why they agreed strongly with the statements that they gave the score "+3" and why they disagreed strongly with the statements that they gave the score "-3". This supported a valid and fast interpretation of factors in the last step of Q methodology (cf. Steelman and Maguire 1999). Furthermore, we asked the respondents after each sort whether they encountered technical problems or problems with understanding the statements, and whether they missed any statement. Performing the sorting task and answering the questions cost the respondents about 15-30 minutes.

5. Analysis and interpretation

We used the PQMethod software[4] to support analysis of the obtained Q sorts (individual scoring patterns) using factor analysis. Factor analysis is a statistical data reduction technique that is used to explain as much of the variability among the observed Q sorts as possible in terms of a few unobserved scoring patterns, which can be called "shared perspectives" or in more technical terms "factors"[5]. First, PQMethod used principal component analysis to calculate the eight factors with the highest explanatory value, as well as the ratio of the total variance between the Q sorts that each factor explained. Then, we chose the number of factors to be included in the analysis. Only factors that explained more of the total variance than a single Q sort could be included (in other words, the Eigenvalue of each factor should be larger than 1; Donner 2001). Other criteria for the choice of the number of factors were the number of Q sorts that determined each factor, and the number and internal logic of the statements that distinguished each factor from the

[3] In the Delft case, we also performed two face-to-face Q sorting interviews.
[4] Freely available at http://www.qmethod.org/.
[5] Both terms are used interchangeably

other factors. Thus, in order to choose the number of factors, we had to repeatedly analyze the content of sets of different numbers of factors.

After we chose an appropriate number of factors for further analysis, PQMethod clarified the structure of the factors by objectively maximizing variance between each of them using Varimax rotation[6]. PQMethod also calculated the "factor loadings", which express the correlation between each individual Q sorts and each factor. Subsequently, we selected which individual Q sorts would define each factor. We selected all Q sorts with a statistically significant and clean loading on a specific factor. A factor loading is significant when it exceeds a threshold value that is based on the number of statements in the Q set[7] and it is clean when it exceeds the loading on other factors with a certain threshold value[8]. In Figure 4.1 for example, Q sort 1 does not have a significant loading on factor A or B. Q sorts 2 and 3 have a significant and clean loading on factor A, and Q sorts 5 and 6 on factor B. Q sort 4 has a significant loading on both factors, but the loading is not clean. Therefore Q sort 4 is selected as a defining Q sort for neither factor A nor factor B.

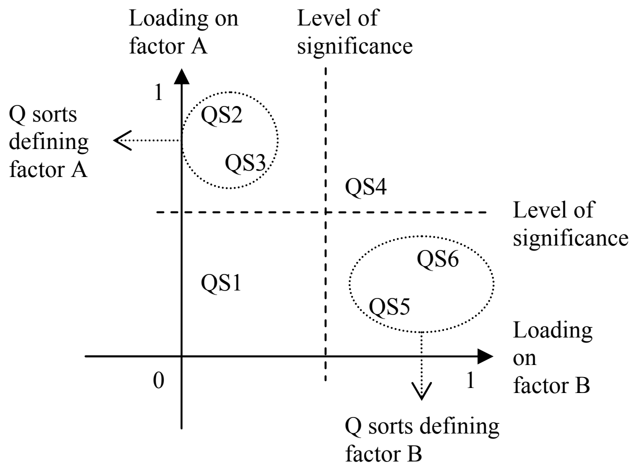

Figure 4.1. Factor loadings and the Q sorts defining a factor (QS = Q sort)

After we selected the defining Q sorts, PQMethod calculated ultimate factor scores and factor Q sort values for each statement under each factor. Ultimate factor scores are the average scores of the respondents defining that factor, weighted by their factor loadings[9]. To obtain factor Q sort values, the statements are ranked according to their ultimate factor score and integer values from -3 to +3 are assigned to them, according to the same distribution that is used for the individual Q sorting (e.g., the score distribution in Table 4.2).

In addition, we calculated the standard error for the ultimate factor scores (SE_{fs}), using Equations 4.1-4.3 (Brown 1980, p. 264 and 297). Equation 4.1 calculates the standard deviation of a Q sort score (s_x), based on the scores in the distribution (x_i), the frequency with which each score occurs (f_i) and the total number of statements in the Q set (N). Equation 4.2 calculates the factor reliability ($r_{xx,factor}$). The factor reliability is based on the test-retest reliability of a Q sort (r_{xx}) and the number of defining respondents for the factor (M). The test-retest reliability is the expected correlation of two Q sorts repeated by the same person, reflecting the variability in sorting without cognitive learning. We adopted the conservative figure of 0.8 for the test-retest reliability (Frank 1956 in Brown 1980, p. 297). Finally, Equation 4.3 calculates the

[6] Alternatively, more subjective manual factor rotation can be used when the analyst aims to confirm a certain prior idea or theory (van Exel and de Graaf 2005).

[7] The limit for the statistical significance of a factor loading is calculated as the multiplier for the desired level of statistical significance (2.58 for $p < 0.01$) divided by the square root of the number of statements N (van Exel and de Graaf 2005).

[8] Since at first we could not find threshold values for a clean loading in literature, we decided to use fixed values of 0.1 (Lower Rhine case) and 0.12 (Delft case, where the Q set was smaller). Later we found the appropriate formula for calculating whether two factors loadings significantly differ (Brown 1980, p. 301) and found out that the values we used reflect a confidence interval of only circa 65%.

[9] Individual Q sorts with high factor loading get a bigger relative weight.

standard error for the ultimate factor scores (SE_{fs}). This number indicates the probable range within which the true factor scores are likely to be located. Since errors may be assumed to fall within a normal distribution, the true factor score will deviate at maximum 1.96 times the standard error from the ultimate factor score with a confidence level of 95% (two-tailed z-test with $p < 0.05$).

$$s_x = \sqrt{\frac{\sum_{i=1}^{N} f_i {x_i}^2}{N}} \tag{4.1}$$

$$r_{xx,factor} = \frac{Mr_{xx}}{1 + (M-1)r_{xx}} \tag{4.2}$$

$$SE_{fs} = s_x \sqrt{1 - r_{xx,factor}} \tag{4.3}$$

PQMethod produced several outputs that were useful for further analysis. Essential were the contention statements, for which factor scores differed significantly between at least two factors, and consensus statements, for which factor scores did not differ significantly between any pair of factors. Based on these outputs, we interpreted the logic of each factor and named each factor. We combined Q methodology with argumentation theory (Fischer 1995; Hoppe and Peterse 1998; Toulmin 1958; see Figure 3.1), in order to recognize the internal logic of each factor. We reconstructed the argumentation structure for each factor, using only the highest and lowest scoring statements in that factor[10]. After analyzing each factor separately, we analyzed the main points of agreement and disagreement between the factors in order to outline areas of consensus and conflict (cf. Steelman and Maguire 1999). We identified conflicting values and interests and conflicting technical knowledge.

Finally, we disseminated and used the results of the analysis in order to 1) support the set up of the research and the collaborative process in each case, by identifying controversial issues that should be discussed (cf. Focht 2002), 2) promote reflection among the collaborating stakeholders in the cases and increase awareness of similarities and differences between their perspectives, and 3) raise awareness among a broad audience[11].

Assessing changes in perspectives over time

We used common statistical analysis to assess the differences between the individual perspectives before and after the collaborative process. More specifically, we analyzed the degree and content of change in overall perspectives, the degree of change in the direction of the research results and the degree of change in the correlation between the perspectives of multiple individuals over time.

First, we calculated the correlation between the Q sorts of an individual before (X_1) and after the collaborative process (X_2). Equation 4.4 (Brown 1980, p. 272) calculates Pearson's correlation coefficient for forced distribution data between two Q sorts X and Y (r_{xy}), in which x and y are deviation scores around the mean of the scores in Q sorts X and Y (which is 0). A correlation between X_1 and X_2 (r_{x1x2}) that is significantly lower than the test-retest reliability r_{xx} indicates a significant change in the perspective. To assess the significance of the difference between r_{x1x2} and r_{xx}, we transformed both values into Fisher's Z (Equation 4.5; Brown 1980, p. 287) and calculated the standard error SE_{Zr} (Equation 4.6; Brown 1980, p. 287). Then we performed a one-tailed z-test ($p < 0.025$) to test whether the difference between the two Z_r-values was significantly greater than 0 (Equation 4.7).

[10] With factor Q sort values of -3, -2, +2 and +3.

[11] We only used the identified shared perspective in the Lower Rhine case to raise awareness among a broad audience, by presenting them at a conference (Raadgever et al. 2007) and in a scientific journal (Raadgever, Mostert and van de Giesen 2008).

$$r_{xy} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{(\sum_{i=1}^{N} x_i^2)(\sum_{i=1}^{N} y_i^2)}} \tag{4.4}$$

$$Z_r = 1.15129 \log_{10}\left(\frac{1+r}{1-r}\right) \tag{4.5}$$

$$SE_{Zr} = \frac{1}{\sqrt{N-3}} \tag{4.6}$$

$$z = \frac{Z_{rx1x2} - Z_{rxx}}{SE_{Zr}} \tag{4.7}$$

Second, we analyzed changes on the level of individual statements. A large difference between an individual's ex ante and ex post score on a statement would indicate that the individual learned about that statement. To be able to assess the significance of the difference scores, we calculated the difference scores D_i (Equations 4.8), as well as the standard deviation SE_D of the difference scores for each individual (Equation 4.9). Then we performed a two-tailed z-test ($p < 0.05$) to test whether each difference score was significantly greater than 0 (Equation 4.8).

$$D_i = X_{2,i} - X_{1,i} \tag{4.8}$$

$$SE_D = \sqrt{\frac{\sum_{i=1}^{N}(D_i - D_{i,average})^2}{N-1}} \tag{4.9}$$

$$z = \frac{D_i}{SE_D} \tag{4.10}$$

Third, we identified about which themes the perspectives changed most strongly, based on the ratio of significant changes in the scores of statements about specific themes, averaged over groups of respondents. This way, we analyzed whether the respondents' perspectives on problems, goals and management strategies changed over time (cf. Figure 3.1), and whether specific groups learned more about specific themes than other groups.

Fourth, we calculated changes in the average scores of groups of respondents on specific statements and calculated whether these changes were significant using a sample-sample t-test. We used Equation 4.11 to calculate the standard error of the difference score of a group of M respondents on one statement. Then we performed a two-tailed t-test ($p < 0.05$) with M-1 degrees of freedom to test whether the difference between the ex ante sample and the ex post sample of scores on a specific statement was significantly greater than 0 (Equation 4.12). The results of this analysis provide an overview of the content and direction of change in specific groups of respondents. The results should be treated with some care, however, because Q methodology measures entire perspectives, in which the statement scores of individuals are mutually related.

$$SE_D = \sqrt{\frac{\sum_{j=1}^{M}(D_j - D_{j,average})^2}{M-1}} \tag{4.11}$$

$$t = \frac{D_{j,average}}{SE_D / \sqrt{M}} \tag{4.12}$$

Fifth, we analyzed whether individual perspectives changed in the direction of the presented and discussed research results. We first selected the statements that were clearly supported or rejected by the research results and determined for each of these statements which score would reflect the research results best. When the research results strongly supported the statement, we assigned the score "+2/+3", and when the research results strongly rejected the statement, we assigned the score "-2/-3"[12]. Then, we determined for each respondent to what extent his or her Q sorting scores on the selected statements changed in the direction of the scores that were expected based on the research results. The final measure of (individual) leaning from the research results was obtained by summing up the changes in scores in the direction of the research results and subtracting from this sum the changes in the opposite direction. The total was divided by the sum of the expected changes in scores based on the research results. When the resulting value turned out to be larger than 0.1, we concluded that the individual perspective changed predominantly in the direction of the research results. When it turned out lower than -0.1, we concluded that the individual perspective changed predominantly in the opposite direction. We did not test the statistical significance of the changes.

Finally, we assessed whether individual perspectives converged, towards a greater consensus among groups of respondents. Consensus between multiple Q perspectives, at a specific moment in time, was calculated as the average of the correlation coefficients (r_{xy}) between each pair of individual Q sorts in a specific group of respondents. An increase in the average correlation coefficient indicates an increase in consensus. To test the statistical significance of the change in consensus in a group respondents we calculated the difference $D_{ri,j}$ between the ex ante correlation coefficient r_{i1j1} and the ex post correlation r_{i2j2} for each pair of Q sorts. This resulted in M values for $D_{ri,j}$, for which we calculated the mean value and standard error (cf. Equation 4.11). Then, we performed a t-test with M-1 degrees of freedom to check whether the difference between the ex ante sample and the ex post sample was significantly larger than 0 (two-tailed t-test with $p < 0.05$; cf. Equation 4.12).

References

Brown, S. R. (1980). *Political subjectivity: Application of Q methodology in political science*. Yale University Press, New Haven.

Denzin, N. K., and Lincoln, Y. S. (2000). *Handbook of qualitative research*. Sage, Thousand Oaks.

Donner, J. C. (2001). *Using Q-sorts in participatory processes: An introduction to the methodology*. Social Analysis: Selected Tools and Techniques. The World Bank, Social Development Department, Washington DC.

Dryzek, J. S. (1993). Reconstructive democratic theory. *The American political science review,* 87(1): 48-60.

Eden, C. (1988). Cognitive Mapping: A Review. *European Journal of Operational Research,* 36: 1-13.

Evans, J. S. T. (1988). The Knowledge Elicitation Problem - a Psychological Perspective. *Behaviour & Information Technology,* 7(2): 111-130.

Fischer, F. (1995). *Evaluating public policy*. Nelson-Hall, Chicago.

Focht, W. (2002). Assessment and management of policy conflict in the Illinois River watershed in Oklahoma: an application of Q methodology. *International Journal of Public Administration,* 25(11): 1311-1349.

Frank, G. H. (1956). Note on the reliability of Q-sort data. *Psychological reports,* 2(182).

Hoppe, R., and Peterse, A. (1998). *Bouwstenen voor argumentatieve beleidsanalyse*. Elsevier Bedrijfsinformatie, The Hague.

McKeown, B., and Thomas, D. (1988). *Q methodology*. Sage university papers. Quantitative applications in the social sciences 66. Sage, Newbury Park.

Pahl-Wostl, C., and Hare, M. (2004). Processes of social learning in integrated resources management. *Journal of Community & Applied Social Psychology,* 14(3): 193-206.

[12] A more specific judgment was considered unfeasible, since different interpretations are possible of the discussed uncertainties, conditions of exception, and emphasis that the statement received during the workshops. Furthermore, the claims about a statement may be different in different sub group discussions.

Raadgever, G. T., Becker, G., and Mostert, E. (2007). Perspectives on future flood management: discovering patterns of thinking. Pages 186-191 *in* Van de Giesen, N. C., Jun, X., Rosjberg, D., and Fukushima, Y. (Eds.), *Changes in Water Resources Systems: Methodologies to Maintain Water Security and Ensure Integrated Management*. IAHS Press, Wallingford.

Raadgever, G. T., Mostert, E., and van de Giesen, N. C. (2008). Identification of stakeholder perspectives on future flood management in the Rhine basin using Q methodology. *Hydrology and Earth System Sciences,* 12(4): 1097-1109.

Ridder, D., Mostert, E., Wolters, H. A., Cernesson, F., Echavarren, J. M., Enserink, B., Kranz, N., Maestu, J., Maurel, P., Otter, H., Patel, M., Schlussmeier, B., Tabará, D., and Taillieu, T. (2005). *Learning together to manage together - Improving participation in water management*. Druckhaus Bergmann, Osnabrück.

Rugg, G., and McGeorge, P. (1997). The sorting techniques: A tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems,* 14(2): 80-93.

Steelman, T. A., and Maguire, L. A. (1999). Understanding participant perspectives: Q-methodology in National Forest Management. *Journal of Policy Analysis and Management,* 18(3): 361-388.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press, Cambridge.

van Exel, N. J. A., and de Graaf, G. (2005). *Q Methodology: A sneak preview*. Retrieved from http://www.qmethodology.net/.

van Tubergen, G. N., and Olins, R. A. (1979). Mail vs personal interview administration for Q sorts: a comparative study. *Operant subjectivity,* 2(2): 51-59.