Investigating change in subjectivity: the analysis of Q-sorts in longitudinal research

Article

Published Version

Morea, N. ORCID: https://orcid.org/0000-0003-0623-3078 (2022) Investigating change in subjectivity: the analysis of Q-sorts in longitudinal research. Research Methods in Applied Linguistics, 1 (3). 100025. ISSN 2772-7661 doi: 10.1016/j.rmal.2022.100025 Available at https://centaur.reading.ac.uk/122132/

To link to this article DOI: http://dx.doi.org/10.1016/j.rmal.2022.100025

Publisher: Elsevier

Research Methods in Applied Linguistics 1 (2022) 100025

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Investigating change in subjectivity: The analysis of Q-sorts in longitudinal research

Nicola Morea*

Faculty of Education, University of Cambridge, 184 Hills Rd, Cambridge CB2 8PQ, UK

ARTICLE INFO

Keywords:
Q methodology
Systematic review
Q-sort
Mixed-methods
Longitudinal

ABSTRACT

A growing number of studies in applied linguistics have used Q methodology to systematically explore language learners' and teachers' subjectivity (e.g., opinions, beliefs, identity, emotions). However, very few studies have used Q methodology and its technique to investigate change over time or after an intervention, partly because of a lack of guidelines on how to analyse Q-sorts completed by the same subjects at different time points. This article aims to provide guidance on how paired Q-sorts can be analysed and to discuss the opportunities and challenges in using the Q-sort method within repeated-measures research designs. To achieve this aim, a systematic review of longitudinal studies (observational and experimental) across research fields was conducted. Twenty-five studies that met the inclusion criteria were selected and scrutinised. Although the review revealed a high degree of heterogeneity in the approaches used to analyse paired Q-sorts, these could be grouped into three broad categories, labelled "Q-factor analysis", "descriptive statistics" and "inferential tests", which are summarised and critically discussed. Based on the findings of the review and in line with the principle of holistic inquiry characterising Q methodology, a mixed-methods analytical approach is then proposed and exemplified using a dataset from a quasi-experimental study on pre-service teachers' beliefs about multilingualism.

Q methodology (henceforth Q) is a series of procedures designed to systematically study human subjectivity (e.g., opinions, beliefs, identity, emotions). Introduced by physicist and psychologist William Stephenson in the 1930s, Q has been adopted to investigate people's perspectives in various research fields, including political science, medicine, education, psychology and, more recently, applied linguistics. Within applied linguistics, Q has been used to explore language learners' motivation (Caruso & Fraschini, 2021; Zheng et al., 2019, 2020) and pre-service and in-service teachers' subjectivity and emotions, such as language teachers' anxiety (Fraschini & Park, 2021), teachers' beliefs about multilingualism (Lundberg, 2019a, 2019b) and teaching competencies (Irie et al., 2018). Additionally, the last few years have witnessed a growing interest in Q among applied linguists concerned with exploring subjectivity from a Complex Dynamic System Theory (CDST) perspective. To the author's knowledge, the first published study reporting the use of Q within a CDST framework was Irie and Ryan's (2015) exploration of the dynamics of change in language learners' self-concepts. MacIntyre et al. (2017) later referred to Irie and Ryan's research to demonstrate the potential of Q for the investigation of dynamic systems, and Zheng et al. (2020) more recently used Q to track the dynamics of change in multilingual motivation among a group of Chinese university students involved in L2 and L3 learning.

* Corresponding author.
E-mail address: nm613@cam.ac.uk

https://doi.org/10.1016/j.rmal.2022.100025
Received 15 March 2022; Received in revised form 4 July 2022; Accepted 4 July 2022

Although Q is becoming increasingly used in applied-linguistics research exploring subjectivity cross-sectionally, the two longitudinal studies by Irie and Ryan (2015) and Zheng et al. (2020) seem to be the only examples of applied linguistics research that used Q to explore subjectivity over time. Similarly, so far virtually no study seems to have adopted Q within an experimental or quasi-experimental design, and current (quasi-)experimental studies still rely on traditional quantitative methods to assess the effects of interventions on participants' perspectives. This methodological gap was acknowledged by Schroedler and Fischer (2020) in their investigation into the development of pre-service teachers' beliefs about multilingualism after taking part in a teacher training course; in the conclusion, the researchers reflect on the limitations of Likert items to explore the complexities of belief change over time and propose the implementation of the Q-sort method in future studies, as it "could potentially shed light on the issues that remain unclear from the data presented here" (p. 20).

One of the possible reasons why Q has rarely been used in studies adopting repeated measures is that there is currently no consensus on how to conduct within-subjects analysis of Q-sort data. Whilst the procedures for using Q cross-sectionally are well-established, the use of Q in repeated-measures research represents an unusual design across all fields. The overarching aim of this article is thus to provide a reference on how paired Q-sorts can be analysed. This objective is achieved in two ways: firstly, by conducting a systematic review of longitudinal studies to illustrate how researchers have analysed Q-sorts within subjects, and, secondly, by proposing a mixed approach which draws on the findings of the systematic review and aligns with the principle of holistic inquiry characterising Q methodology. Accordingly, the rest of the article is structured into three sections: Section 2 provides an introduction to Q and the Q-sort method; Section 3 reports and discusses the findings of the systematic review, and Section 4 describes and exemplifies the proposed analytical approach.

Introducing Q methodology

Conducting a Q-methodological study: Data collection and analysis

The sorting activity and the creation of a Q-set

At the core of Q methodology is a sorting activity. Participants are given a number of statements (typically 30-50, Lundberg et al., 2020) on the topic of inquiry (e.g., British teachers' opinions on the perceived benefits and disadvantages of being multilingual), and they are asked to order the statements in a specific way, for example from most agree to most disagree. This list of statements is called a *Q-set*. Each participant is also provided with a grid on which the Q-set statements will be placed (Fig. 1). The grid is made of numbered columns, ranging, for example, from -4 to +4 (but values vary based on the number of statements). Each participant will place those statements with which they most disagree under the -4 column and those statements with which they most agree under the +4 column. Participants will then rank-order the remaining statements on the grid based on their viewpoints. Figs. 1 and 2 show, respectively, an empty Q-sort grid and the same grid populated with statements after a sorting activity. The Q-sort grid is typically shaped as a (quasi-)normal distribution, and the number of statements that can be placed under each column is pre-determined. Therefore, fewer statements can be placed at the two extremes of the grid and more statements can be placed towards the centre of the grid. This 'forced distribution' encourages participants to carefully consider each statement both individually and in relation to the others. The result of the sorting activity is a grid populated with all the statements ordered in a way that reflects the participant's viewpoints (Fig. 2); this grid is a participant's *Q-sort*.

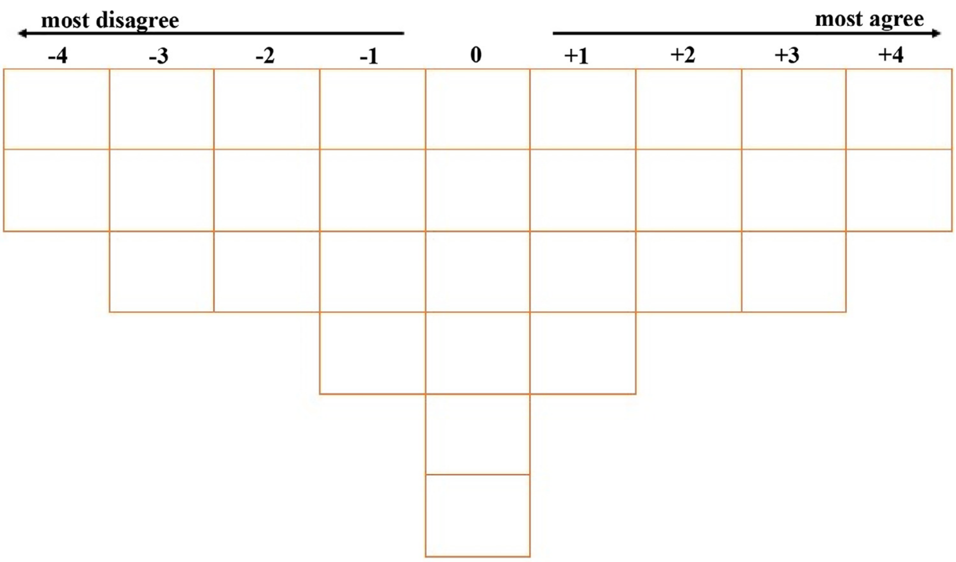

Fig. 1. An Empty Q-Sort grid with space for 30 statements

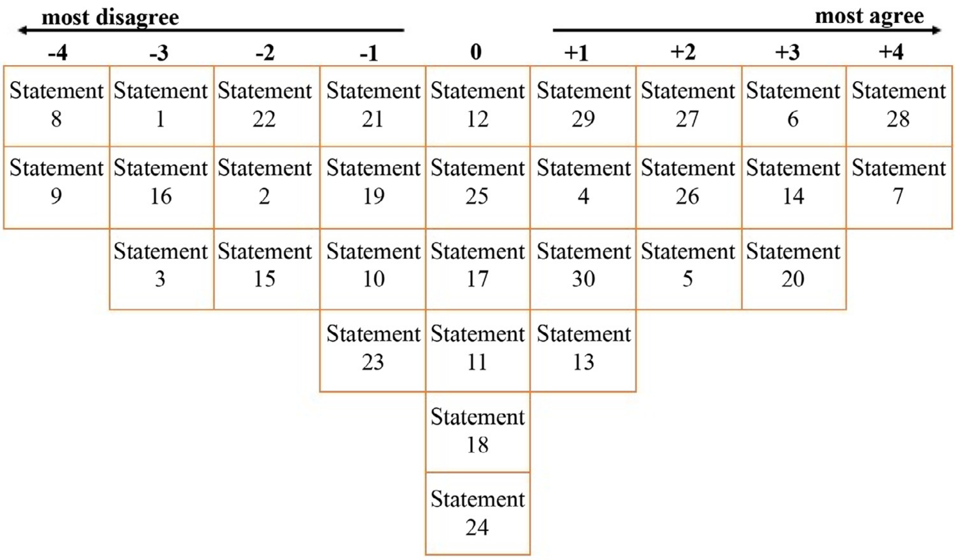

Fig. 2. A complete Q-Sort made of 30 statements
Note. Statement are replaced by statement numbers due to space limitations.

The creation of a Q-set is a process that distinguishes Q from other methodologies. The statements that are included in a Q-set originate from a *concourse*. A concourse is the totality of positions taken on the topic of inquiry (Brown, 1993); if we take the topic of British teachers' opinions on the perceived advantages and disadvantages of being multilingual as an example, the concourse may consist of any advantage and disadvantage that has ever been attributed to multilingualism in the education context in the UK. The researcher would extract a large number of statements from this concourse by examining different sources, such as academic and grey literature, and by interviewing teachers to elicit their viewpoints on the topic. This list of statements is then reduced until the researcher is satisfied that the statements retained (i) cover the main positions found in the concourse and (ii) are balanced, as not to be biased towards specific opinions (Watts & Stenner, 2012). Approaches to validate the Q-set used in the literature include having the Q-set scrutinised by a panel of experts and conducting pilot studies (Lundberg et al., 2020). An example of a Q-set can be found in the supplementary material.

Q-factor analysis

After the sorting activity, the Q-sort data are uploaded on a software specifically designed for conducting Q-factor analysis of Q-sorts. The software first performs a by-person correlation of participants' Q-sorts; in other words, a subject's Q-sort is correlated with the Q-sort of all the other subjects, resulting in a matrix of correlation coefficients indicating how similar, or different, each pair of Q-sorts is. It is important to highlight that these correlations (and the subsequent factor analysis) are conducted on participants' Q-sorts *as a whole*, and not on individual statements. Since, in Q, a Q-sort is considered the manifestation of a person's subjectivity in relation to the topic of inquiry, the correlation coefficients can be interpreted as the degree of similarity (or difference) between participants' viewpoints. The correlation matrix represents the starting point for conducting a Q-factor analysis of participants' Q-sorts. After the researcher selects a preferred factor-extraction method, the software extracts a number of factors from the sample of Q-sorts and produces a scree plot. The researcher then decides on a factor solution (i.e., how many factors to retain from those extracted) based on certain criteria (e.g., factor eigenvalues, explained variance of the factor solution). Typically, 2-4 factors are retained in a Q study (Lundberg et al., 2020), and the retained factors are then rotated to facilitate factor interpretation.

The extent to which each subject's Q-sort loads on each rotated factor is expressed through a factor loading, namely a correlation coefficient between a subject's Q-sort and the factor. The researcher selects those Q-sorts whose factor loading on a factor is statistically significant (usually at the .05 or .01 level). The selection of those Q-sorts that significantly load on each factor determines which Q-sorts the software will use to calculate factor scores. Whilst it is expected that a participant's Q-sort will load *significantly* on one Q-factor only, it is not uncommon to see confounded Q-sorts, namely Q-sorts that significantly load on more than one factor. These Q-sorts tend not to be selected for factor interpretation (Watts & Stenner, 2012).

Factor interpretation

Once the Q-sorts significantly loading on each extracted factor have been selected, it is possible to start interpreting the factors. For every retained factor, the software generates a factor array, namely a *factor-defining* Q-sort in which each statement of the Q-set is placed on the original grid based on its factor score. Taking the grid in Fig. 1 as an example, those statements with the two lowest and highest factor scores are placed in the +4 and -4 columns, the successive three in the +/- 3 column, and so on until all statements have been placed on the grid. Therefore, a factor array can be considered an archetypical Q-sort that displays the shared configuration of viewpoints held by a group of participants as expressed through a Q-factor.

To understand the characteristics of each factor (i.e., which configuration of viewpoints each factor expresses), the researcher carefully examines the factor arrays, with a particular focus on *distinguishing* and *consensus* statements. A factor's distinguishing statements are statements whose factor scores are statistically significantly different between factors (i.e., statements that participants significantly loading on a factor sorted differently from other participants) (Coogan and Herrington, 2011). Consensus statements are those statements whose factor scores are similar across factors and thus represent shared opinions within the sample. The interpretation and description of Q-factors is not, however, limited to an analysis and comparison of the factor arrays. After the sorting activity, additional data are collected through post-sorting interviews with research participants, during which participants are invited to explain the reasons behind their sorting decisions. As most participants will be associated with a factor, these qualitative data provide important information to guide factor interpretation and, at the same time, understand participants' viewpoints in relation to their individual background and context.

Comparing Q and 'R methodology'

Q was proposed in the 1930s in opposition to what Stephenson, the creator of Q methodology, referred to as 'R methodology', namely the set of statistical methods conventionally used in psychology to measure individual traits through scales. A typical example of R methodology would be the use and analysis of Likert scales, which are made of a number of Likert items whose design is based on the researcher's own hypothesis or theory around the nature of the construct they aim to measure. The results obtained from such scales, and the reliance on large numbers of subjects to estimate population parameters, were considered by prominent Q-methodologists as "merely artifacts reflecting the way in which the data were collected", so that "the individual's independent point of view, in effect, is considered to be dependent on the prior meaning of the scale" (Brown, 1980, p. 4). Q was designed with the aim of studying an individual's viewpoints without constrictions from the investigator. Stephenson defined Q as "a methodology for the single case" (p. 12), proposing the methodology in opposition to the use of large samples and the examination of individual differences. As such, Q is fundamentally an exploratory methodology. However, this does not mean that results from a Q-study cannot be generalised. Although single-case Q studies exist, most Q research involves collecting Q-sorts from a sample of participants (Watts & Stenner, 2012). In investigations using multiple-participant designs, Q allows the researcher to reveal and describe shared viewpoints that *exist* in a sample and, by extension, in the population of reference (Stephenson, 1953). Since Q is not concerned with estimating how frequent certain viewpoints are in a population, but rather with uncovering viewpoints that exist in that population, a Q-method study does not require a large number of participants to achieve its objectives (Watts & Stenner, 2012). For example, in a review of 74 Q studies in the field of education, Lundberg et al. (2020) found that the mean number of participants (or *P-set*) was 37.

Potential of Q for applied-linguistics research

The potential of Q for exploring subjectivity is that it allows a holistic investigation of human perspectives. This is particularly evident when comparing the Q-sort method to traditional quantitative methods, such as Likert items and scales: on the one hand, the sorting activity requires participants to consider and compare all the statements in the Q-set and rank them in a way that best reflects their overall point of view (Watts & Stenner, 2012); on the other hand, the interpretation of the extracted factors requires an examination of the *whole* configuration of statements that characterises each factor, together with the incorporation of participants' post-sorting reflections, as to account for the contextual reasons behind participants' factor association. As a result, however, a Q data collection process is more time-consuming and cognitively demanding for the research participant than a questionnaire, a potential drawback that becomes even more prominent in designs involving multiple data collections. Researchers should also be aware that whilst Q can be used to reveal shared viewpoints that exist in a population, it does not allow to make any generalisations on how widespread the viewpoints uncovered are in that population, nor to statistically examine associations between factor alignment and participant characteristics (Watts & Stenner, 2012).

Whilst providing a methodological and analytical framework for future applied linguistics research investigating subjectivity around language-related phenomena cross-sectionally (e.g., language learners' and teachers' motivation, linguistic identity, beliefs), Q can also be used within longitudinal research designs examining change in subjectivity over time. However, longitudinal Q studies are much less frequent than cross-sectional studies (Lundberg et al., 2020), and there does not seem to be any established procedure to analyse Q-sorts collected from the same participants at multiple time points. Accordingly, the next section presents and discusses what analytical approaches have been used across research fields to analyse paired Q-sorts.

A systematic review of longitudinal studies using the Q-sort method

Method

A systematic review of the literature was conducted by the author between November 2021 and February 2022. The stages of the review are displayed in Fig. 3 using the PRISMA flow diagram for systematic reviews (Page et al., 2021). The following databases were consulted on 13th January 2022 for the identification and selection of studies: Scopus, Web of Science and all the accessible

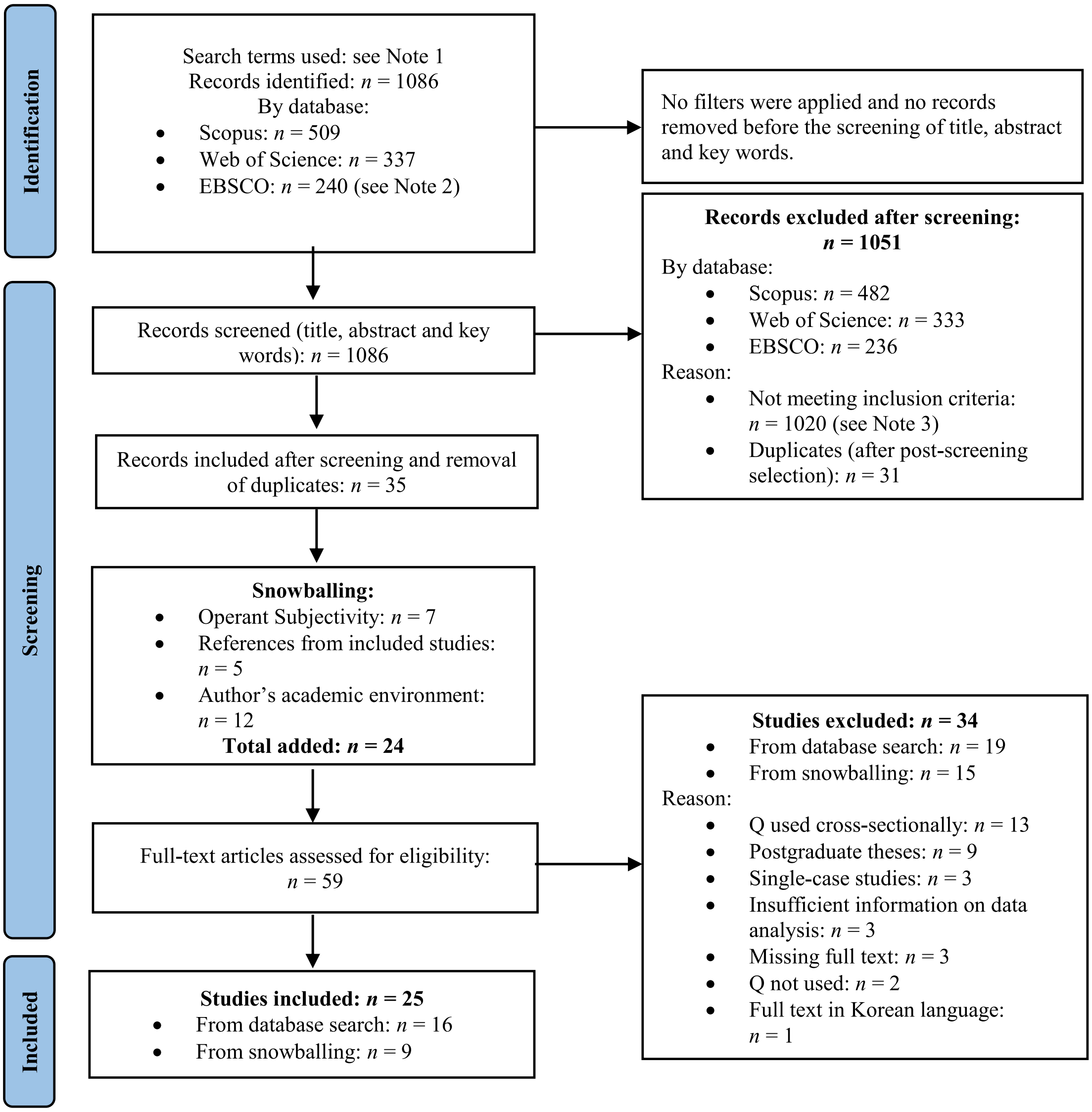

Fig. 3. Systematic review phases
Note 1. String used: ("q method" OR "q-method") AND (longitudinal OR "prospective study" OR "cohort study" OR "follow-up study" OR "panel study" OR "before and after" OR experiment OR quasiexperiment OR quasi-experiment OR pre-test OR pretest OR post-test OR posttest OR "pretest-posttest" OR "pre-post" OR prepost OR intervention OR "time points" OR "over time"). Note 2. Due to the limited number of search terms that could be included in a single search on the EBSCO platform, a series of simple searches was conducted using the string: q method AND (each search term of the second part of the string). The reported number of records is the total number of results obtained. Note 3. The reason for the high number of exclusions at this stage is due to a large proportion of studies that used Q cross-sectionally but included in the abstract one of the search terms from the above string.

databases within the EBSCO Host research platform (see Appendix for the complete list). Considering the relative paucity of studies that have used Q in repeated-measures designs, no restrictions were imposed regarding the year of publication or research field.

Similarly, the decision to review both observational and (quasi-)experimental studies was justified by the need to maximise the number of selected studies and by the fact that both designs typically involve collecting data from the same participants at two or more time points. However, there are important differences between the two designs: whilst (quasi-)experimental studies introduce some form of intervention with the aim of assessing its effects on participants, non-experimental longitudinal studies are observational and tend to gather data over an extended period of time (Cohen et al., 2007).

To maximise the exclusion of studies that used the Q-sort method cross-sectionally, the search string included "Q method" as the primary search term, followed by a list of terms specific to longitudinal designs (see Note 1 under Fig. 3 for the full string). The search terms were selected after consulting Cohen et al. (2007) manual of research methods in educational research, other systematic reviews of longitudinal research (e.g., Chaiton et al., 2009) and studies aligning with the selection criteria known to the author. In defining the search terms, the aim was to strike a balance between comprehensiveness and precision (Bramer et al., 2018); for example, it was decided not to include the search term "time" due to the high number of irrelevant records that would have ensued, and to use specific collocations ("over time", "time points", "before and after") instead.

The database search resulted in a total of 1086 records. A screening of titles, abstracts and key words was then conducted against the following inclusion criteria: (i) the study uses the Q-sort method, and (ii) the study compares Q-sorts obtained from the same participants at two or more time points. After the initial screening and during the final selection phase (full-text records assessed: $n = 59$), three additional sub-requirements within criterion two were added to increase the selection precision:

2.1 The Q-set and condition of instruction must be the same across time points.
2.2 The study must examine one or more groups of subjects instead of individual participants (e.g., exclusion of single-case designs).
2.3 The author(s) must provide sufficient information to understand the analytical procedures used.

Finally, only completed and published research was selected; postgraduate theses, conference presentation summaries and research proposals were excluded. A full examination of the records selected after screening resulted in 16 studies that met the inclusion criteria.

Following the approach adopted by Lundberg et al. (2020) in their systematic review of Q studies in the field of education, a snowballing approach was used to maximise the inclusion of eligible records. Specifically, nine more studies were added from: (i) studies referenced in the selected records; (ii) studies known to the author or their academic environment; (iii) studies selected after a screening of all the issues of Operant Subjectivity, the peer-reviewed, non-indexed international journal of Q methodology. The inclusion of the above records resulted in a total of 25 studies.

As indicated in the inclusion criteria, studies were included based on whether Q-sorts were generated and analysed over time. Consequently, not all the included studies are necessarily *Q-methodological* studies, that is, studies that adopted all the procedures of Q (including, for example, creating a Q-set from a concourse of reference or exploring individual subjectivities through post-sort interviews). As one of the aims of this article is to review how paired Q-sorts have been analysed, and in consideration of the relative scarcity of studies that analysed Q-sorts over time, the use of the Q-sort method was considered a pre-condition for the inclusion of a study in the review rather than a study's alignment with all the principles of Q.

Characteristics of selected studies

Context and research designs

The 25 selected studies spanned from 1970 to 2022 and belonged to various research fields. Education was the most represented field ($n = 10$), and the majority of studies belonging to other disciplines involved an element of teaching or learning. Other fields included: political sciences ($n = 9$), applied linguistics ($n = 2$), medical sciences ($n = 2$), land economy ($n = 1$) and methodology ($n = 1$). Most studies were conducted in the United States ($n = 16$), three in Australia, three in Europe (England, Germany and the Netherlands), one in Japan and one in China. Most studies ($n = 20$, 80%) were quasi-experimental or experimental, whereas five were observational. Of the former group, only one study can be considered a true experiment, whereas the remaining ($n = 19$) either did not include a control group ($n = 17$) or, if they did, group assignation was not randomised ($n = 2$).[1] The predominance of non-probability sampling is not surprising: on the one hand, the randomisation of subjects in educational settings may often be unfeasible or even unethical (Cohen et al., 2007); on the other hand, Q studies have traditionally adopted purposive sampling techniques given the smaller sample size required and the time-consuming sorting activity (Watts & Stenner, 2012). Among (quasi-)experimental studies, the sample sizes of the experimental groups were relatively small: excluding an outlier study with a sample size of 141, the sample of the remaining studies ranged from 11 to 54, $M = 26.82$, $SD = 12.99$, $Mdn = 28.5$ (outlier included). Observational studies showed a similar trend: with the exception of one study with a sample size varying between 65 and 97 (depending on the number observations at each of the four time points), the sample size of the other studies was 15, 19, 34 and 36/37 (depending on the time point). Eight studies explicitly reported issues of attrition, namely having more data at baseline than at subsequent time points due to participant withdrawal, whereas it is impossible to ascertain how many of the 17 remaining studies did not experience attrition or did not report it. Finally, 20 studies involved the administration of Q-sorts at two time points, three studies at three time points (two quasi-experimental and one observational) and two studies at four time points (one quasi-experimental and one observational).

Q-specific characteristics

This section reports characteristics of the selected studies that are specific to Q-factor analysis. All studies except one conducted a Q-factor analysis of the Q-sorts at baseline. The number of statements in the Q-set ranged from 20 to 80, $M = 47.94$, $SD = 13.16$. The most used software for Q analysis was PQ-Method (Schmolck, 2014); other pieces of software used between 2000 and 2022 were KADE (Banasick, 2019), PCQ (Stricklin & Almeida, 2001), QFactor (Akhtar-Danesh, 2018) and QMethod (Zabala, 2014). Regarding factor extraction, nine studies (37.5%) used the centroid method, six (25%) principal component analysis (PCA), whereas nine studies

[1] In line with Cohen et al.'s (2007) discussion of experimental research designs, "quasi-experimental" is used in this article to refer to studies adopting either a non-equivalent control pre-test-post-test design or a one-group, pre-test-post-test design. On the other hand, "experimental" is used to refer to randomised control pre-test-post-test designs.

(37.5%) did not report the method used. With regard to factor rotation, varimax was used in 11 studies (45.83%) and manual rotation[2] in five (20.83%); three studies (12.5%) adopted a mixed approach, using both varimax and manual rotation, whereas five studies (20.83%) did not report this information.

Eleven studies (45.83%) did not mention the criteria used to select the number of extracted factors to retain prior to rotation. Among the remaining studies, the criteria used were based on: a minimum number of participants significantly loading on a factor (at least two: $n = 3$, more than two: $n = 1$, cut-off number not specified: $n = 2$); factor eigenvalues > 1 ($n = 5$); overall explained variance of the factor solution ($n = 2$, but no cut-off point specified); *Humphrey's rule*[3] ($n = 2$); whether the factor could be clearly interpreted ($n = 2$); the results of a scree test ($n = 1$). Most studies adopted one of these criteria, whereas three studies used two or more. Overall, the number of factors extracted at baseline (including studies that factor-analysed Q-sorts from different time points together) ranged from 1 to 6, $M = 3.42$, $SD = 1.41$, $Mdn = 3.5$, mode = 4. Finally, only four studies mentioned the criteria used for selecting the Q-sorts that would define each factor. This is an important decision in Q studies that closely links with how to treat confounded Q-sorts (i.e., Q-sorts significantly loading on more than one factor). On the one hand, the inclusion of confounded Q-sorts to define a factor may lead to higher factor intercorrelations. On the other hand, this effect may be limited if the confounded Q-sorts have a strong primary factor loading and a much weaker secondary loading. Among the four studies that acknowledged this issue, two nonetheless included all confounded Q-sorts, whereas the other two studies only included confounded Q-sorts if the second significant loading was "weak". Whilst one study did not specify any cut-off value to define a weak secondary loading, Zheng et al. (2020) dealt with the issue by increasing the threshold for a significant factor loading (i.e., correlation coefficient) to be selected for factor interpretation (from .38, the minimum value for a factor loading to be significant at the .01 level, to .42).

Analytical procedures used to compare Q-sorts over time

A review of the analytical procedures used to compare paired Q-sorts revealed a variety of approaches. These included both Q-specific analyses (i.e., Q-factor analysis) and traditional statistical approaches (e.g., analysis of variance, or ANOVA), henceforth referred to as *R-based* analyses in line with Q-methodology conventions. These analytical procedures were grouped into three categories, which are discussed below.

Category 1: Q-factor analysis

Most studies ($n = 20$, 80%) analysed paired Q-sorts using Q-factor analysis, either by factor analysing all Q-sorts together ($n = 9$) or by conducting separate Q-factor analyses at each time point ($n = 11$). The rationale for choosing the former approach may be that it produces a single factor solution accounting for all Q-sorts, which allows to easily assess how many participants remained aligned with their original factor. However, as discussed by Akthar-Danesh and Wingren (2020), such analysis does not account for the inter-correlation between Q-sorts from matched pairs, as it includes between- and within-subject Q-sorts in the same analysis. For this reason, this approach may also not be suitable to detect the emergence of new factors over time.

Other studies conducted a Q-factor analysis at each data collection point. As a result, a number of factors were obtained from each analysis, raising the question of how these factors can be compared. Most studies did not provide a statistical measure of similarity between baseline and successive factors, and instead subjectively interpreted paired factors by comparing their factor scores (Zheng et al., 2020), distinguishing statements (Tornwall & McDaniel, 2022) and statements with the highest and lowest factor scores between factors (Irie & Ryan, 2015). A smaller number of studies conducted a Pearson's correlation between pairs of factors.[4] The correlation coefficient thus obtained provided a measure of how close or different each pair of factors were, helping to assess whether two factors extracted at different time points represented the same or different configurations of viewpoints. It is, however, important to note that the factors obtained from a Q-factor analysis are shaped by decisions regarding the extraction and rotation methods, the criteria for choosing a factor solution and the selection of defining Q-sorts. To compare factors extracted from different Q-analyses, it is therefore essential that a set of criteria is decided a priori by the researcher and consistently applied in every subsequent Q-factor analysis.

Category 2: Descriptive statistics

This category includes analytical approaches that used the factors and factor loadings obtained at time one as a baseline for the comparison of successive Q-sorts. In contrast with category three, these statistical approaches are descriptive; in other words, they describe change in a sample without assessing whether any difference is statistically significant. Four types of analyses are included in this category.

The first type consists of comparing the proportion of participants associated with each baseline factor at different time points. In studies that conducted a single Q-factor analysis of all Q-sorts, this was easily assessed by looking at the distribution of factor loadings of Q-sorts collected at different time points. In studies that conducted a separate Q-analysis at each data collection point, successive Q-sorts were correlated with the factor array of each baseline factor to calculate factor loadings (see Footnote 4). Then,

[2] Manual rotation, or hand rotation, is a rotation technique where the researcher rotates the factors manually. It is also referred to as "theoretical rotation", as it presupposes some initial hypothesis or theory on the resulting factor composition (Watts and Stenner, 2012).

[3] From Watts and Stenner (2012): according to Humphrey's rule, "a factor is significant if the cross-product of its two significant loadings (ignoring the sign) exceed twice the standard error" (p. 107).

[4] Using the factor arrays, two factors can be correlated in the same way as two Q-sorts (see Brown, 1993, for the formula used).

based on the criterion used to select Q-sorts for factor interpretation (generally a factor loading significant at the .05 or .01 level), the researcher(s) assessed the number and percentage of participants that remained significantly aligned with each baseline factor.

Using a similar approach, a second type of analysis consists of comparing the Q-sorts' mean factor loading on each baseline factor between time points. In other words, the mean factor loading of the baseline Q-sorts on a baseline factor was compared with the mean factor loading of the Q-sorts obtained at time two on the same factor, in order to see if participants' overall alignment with each baseline factor had strengthened or weakened over time. However, no inference test was conducted to assess whether any difference in means between time points was statistically significant.

A third type of analysis (used in both observational and experimental studies) aims at assessing the degree of stability between participants' Q-sorts over time. Firstly, each pair of within-subject Q-sorts (the Q-sorts produced by the same participants at two time points) was correlated; then, the range and mean value of all the resulting correlation coefficients were compared between time points to assess the overall stability of participants' viewpoints over time. In (quasi-)experimental studies involving a control group not receiving any intervention, this approach was also used to assess the reliability of the Q-sort method (Cook et al., 1975).

A fourth approach was proposed and exemplified by Akhtar-Danesh and Wingreen (2020). It involves comparing (i) the position in which the distinguishing statements of a baseline factor appear on its factor array (i.e., the factor's Q-sort) with (ii) the average position of the same statements in the successive Q-sorts of those participants who originally (significantly) loaded on the factor. Compared to other types of analysis, this approach allows the researcher to examine change among sub-groups of participants who aligned with a specific factor at baseline. By only comparing sub-groups of participants with similar initial viewpoints, this method accounts for potential issues of intra-class correlation (Akhtar-Danesh & Wingreen, 2020), but may consequently require a larger sample size.

Category 3: Inferential tests

The third category consists of various inferential tests used to statistically assess time change at group level. Most studies used inferential tests to statistically compare participants' factor loadings on each baseline factor at two or more time points, with the aim of assessing whether participants' alignment with each original factor had significantly increased or decreased over time or after an intervention. The most frequently used tests were paired samples t-tests (Freie, 1997; Niemeyer et al., 2013) and ANOVAs (Brown, 1977; Cook et al., 1975). Regarding the latter, studies tended not to specify the type of ANOVA used, although this could be inferred by the study design. For example, studies comparing a single group at two or more time points may likely have conducted a repeated-measures ANOVA (e.g., Brown, 1977); conversely, a two-way mixed ANOVA would have been appropriate when including a control group (e.g., Cook et al., 1975), in order to assess the interaction between time (within-group difference) and condition (between-group difference) on the dependent variable (normally the factor loadings on each baseline factor[5]). A different ANOVA test was used by Cuppen (2012), who conducted a multivariate analysis of variance (MANOVA) using participants' loadings on the baseline factors as dependent variables and adding participants' association with the six baseline factors as a third independent variable (and second within-group variable). Despite using parametric tests, most studies did not report whether the dataset met the test-specific assumptions. The study by Cuppen (2012) seems to be the only one where a violation of univariate normality was discussed. However, the author did not specify why a multivariate test was preferred in a study with a relatively small sample size (experimental group: $n = 11$; control: $n = 12$) over multiple univariate analysis (see Huberty & Morris, 1989, for a comparison between the two). Finally, a different approach was used by Dennis and Goldberg (1996) and Schwartz et al., and Gilliam (2010), who conducted non-parametric tests of association between categorical variables (chi-square test and binomial test with Cramer's V, respectively) to assess whether there was a significant difference in the proportion of participants aligning with the baseline factors at two time points. Arguably, however, other non-parametric statistical tests specifically designed for within-group comparison of distributions may be more appropriate with longitudinal data, such as the McNemar's (McNemar, 1947) and the Bhapkar's (Bhapkar, 1966) tests.

Discussion of the analytical procedures used

The review revealed heterogeneity of analytical approaches used in studies adopting the Q-sort method in repeated-measures designs. The analytical procedures could nonetheless be broadly distinguished into Q- and R-based approaches, although the two should not be considered mutually exclusive: whilst 10/25 studies only adopted one approach (Q-factor analysis only: $n = 4$; R-based analysis only: $n = 6$), 15 studies used both. Within Q-based approaches, a distinction was drawn between studies that conducted a separate Q-factor analysis at each data collection point and studies that conducted a single Q-analysis of all Q-sorts. This latter approach is not well-suited for studies integrating Q- and R-based analysis, since it incorporates between- and within-subjects Q-sorts in a single analysis. With regard to R-based analyses, whilst inferential tests allow the researcher to evaluate whether any change found in the sample is statistically significant and to calculate effect sizes, it is essential that issues related to sample size, sampling technique and test assumptions be acknowledged and discussed. In this regard, it is striking that the majority of studies that used inferential statistics conducted parametric tests, despite an often relatively small sample size and a lack of information on whether the dataset met the test assumptions.

Q- and R-based analyses allow us to explore changes in participants' subjectivity over time from different angles. Q-factor analysis enables the researcher to track the development of shared opinions over time and to uncover the emergence of new configurations of viewpoints and the disappearance of others within a sample. R-based analyses, on the other hand, allow the researcher to assess

[5] With the exception of Brown (1970), where the correlation coefficients of paired Q-sorts were used instead.

the extent to which participants' subjective views have changed over time. Additionally, although used to investigate change at the group level, the above approaches can also provide a basis for an examination of shifting subjectivities at the individual level. In the next section, a mixed analytical approach informed by the above review is proposed, aimed at holistically investigating change and stability of Q-sorts over time. A real dataset is used to exemplify the analytical process proposed.

A mixed approach for the analysis of paired Q-sorts

Overview

The proposed mixed approach consists of four types of analyses. The first two analyses involve the use of within-subjects inferential tests to compare successive Q-sorts with baseline data. Specifically, the first analysis consists of comparing participants' factor loadings (i.e., correlation coefficients) on each baseline factor between different time points, allowing to assess whether participants' alignment with each initial factor has significantly strengthened or weakened at the group level. The second approach first involves assigning a baseline factor to each participant based on participants' highest factor loading at each time point, then testing for time differences in the distribution of participants by their association with the baseline factors. This allows assessing whether participants tended to remain more closely associated with the same factor over time or whether they tended to "move" to (i.e., more strongly load on) a different baseline factor. The third approach consists of conducting a Q-factor analysis at each time point, extracting new sets of factors that are compared to the baseline factors by correlating their factor arrays. This allows assessing the stability of the original factors over time and the emergence of new ones. Finally, the fourth proposed approach is aimed at investigating the processes of subjectivity change at the individual level by collecting and analysing retrospective qualitative data through post-sorting interviews and other qualitative methods. In the next section, the proposed approach is exemplified using a real dataset from a one-group, quasi-experimental study conducted by the author.

Contextual information on the dataset

The data used in this demonstration comes from a larger study conducted by the author which investigated pre-service teachers' beliefs about multilingualism during an Initial Teacher Education and Training (ITET) programme after participating in a course on multilingualism. Fifty-one primary and secondary pre-service teachers in England (training to teach different subjects) were recruited using a voluntary response sampling approach. In October and November 2020, at the start of their ITET programme, they sorted 30 statements on a pre-defined grid based on their level of agreement (with columns ranging from -4, most disagree, to +4, most agree). The Q-set was adapted from Lundberg (2019a) and included a variety of positions on linguistically inclusive practices at school- and classroom level (see the supplementary material for the statement list). From this group, 16 pre-service teachers volunteered to take part in an intervention aimed at furthering their understanding of multilingualism and linguistically inclusive teaching strategies. In May 2021, after the intervention and towards the end of their ITET programme, they completed a second, identical, Q-sort. The aim was to assess whether participants' perspectives had changed after taking part in the intervention. The overarching research question (RQ) was: *did participants' pedagogical beliefs change after participating in the intervention? If yes, how?* To answer this RQ, four types of analysis were used, which answered four sub-questions, namely:

1. RQ1: Did participants' association with the baseline factors strengthen or weaken after the intervention?
2. RQ2: Did the proportion of participants aligning with each baseline factor change after the intervention?
3. RQ3: What factors emerged from participants' Q-sorts after the intervention? What is the relationship between these factors and the baseline factors?
4. RQ4: What trajectories of belief development did participants experience as they engaged with the intervention?

The four approaches are presented below, preceded by an overview of the factors extracted at baseline.

Data analysis

Q-factor analysis at baseline

Participants' Q-sorts at time one were intercorrelated and factor analysed using the KADE desktop application (Banasick, 2019, downloadable from: https://github.com/shawnbanasick/kade). Principal Component Analysis (PCA) with varimax rotation was used for factor[6] extraction and rotation, resulting in the selection of two factors. Details on the criteria used to determine the factor solution and to flag defining Q-sorts can be found in the supplementary material.

A detailed description of the extracted factors is beyond the scope of this exemplification, so a brief summary of their key characteristics is provided. The two extracted factors were only weakly and non-significantly correlated ($r = 0.317$). Pre-service teachers' aligning with the first factor (F1) tended to consider students' linguistic diversity as a resource and were keen to use linguistically inclusive strategies to incorporate students' home languages in their lessons and give students opportunities to actively use their languages in class. They also tended to believe that students for whom English is an additional language (EAL) should receive support

[6] Although the author understands that the term "component" is more appropriate when using PCA, it was decided to follow the convention of Q literature of using the term "factor" also in the context of PCA.

and have the option to be assessed in their home language(s). On the other hand, prepre-service teachers aligning with factor two (F2) tended to consider students' linguistic diversity as an obstacle to their teaching. They did not believe that the classroom is an appropriate environment for using languages other than the language of instruction, and they believed that any provision for EAL students should solely aim at enabling students to acquire the dominant language.

Thirty-two of the 51 participants significantly loaded on F1, ten on F2, whereas eight Q-sorts significantly loaded on both factors and three did not significantly load on any. Of the sixteen participants who formed the experimental group and who later took part in the intervention, seven significantly loaded on F1, four on F2, whereas four Q-sorts were confounded and one did not significantly load on any factor.

Method 1: Pretest-posttest comparison of participants' factor loadings on baseline factors

To test whether the strength of participants' loadings on each baseline factor differed significantly before and after the intervention, a Wilcoxon signed-rank tests was conducted for each baseline factor. The post-test factor loadings on the baseline factors were calculated by correlating each participant's post-Q-sort with the factor array of each baseline factor (see Footnote 4). Although the dependent variables (factor loadings) are continuous, a non-parametric test was preferred over a paired samples t-test given the presence of outliers and the violation of the normality assumption of the factor loadings on F1 (Shapiro-Wilk's test: $p = 0.39$). Tables 1 and 2 report, respectively, information on the ranks before and after the intervention and the statistics of the Wilcoxon signed-rank tests. As the tables show, there was a positive mean increase in participants' factor loadings on factor F1 (+.097) and a slight decrease in factor loadings on factor F2 (-.042), but none of these differences was statistically significant. These results indicate that participants did not tend to align significantly more strongly or weakly with either of the two baseline factors after the intervention.

Method 2: Pretest-posttest comparison of distribution of participants by highest factor loading on baseline factors

A McNemar's test was conducted to test for significant changes in the distribution of participants by their highest factor loading on either F1 or F2 after the intervention. This test is used to statistically compare dichotomous categorical data collected at two time points from the same sample (Leon, 1998). Table 3 shows the distribution of participants on factors F1 and F2, before (rows) and after (columns) the intervention. Before the intervention, eight participants loaded on F1 more strongly than on F2; after the intervention, the number increased to 14, with only two participants still loading more strongly on F2. No participant who initially loaded on F1 aligned more strongly with F2 after the intervention. A McNemar's test revealed that the change in distribution was statistically significant (p = .031). This finding indicates that participants whose beliefs used to align more strongly with profile F2

Table 1
Comparison of pre- and post-test factor correlations, factors F1 and F2.

Factor	Ranks for post-pre correlations	*n*	Mean rank	Sum of ranks
F1	Negative	7[a]	8.07	56.5
	Positive	9[b]	8.83	79.5
	Ties	0[c]		
	Total	16		
F2	Negative	9[a]	8.72	78.5
	Positive	7[b]	8.21	57.5
	Ties	0[c]		
	Total	16		

Note. [a]Post < Pre
[b] Post > Pre
[c] Post = Pre

Table 2
Wilcoxon signed-rank test statistics, pre- and post-test correlations, factors F1 and F2.

Factor	Mean score			Test statistics		
	Pre-intervention	Post-intervention	Difference	*Z*	*p*	*r*
F1	.484	.581	+0.097	-0.596	.551	-.105
F2	.344	.302	-0.042	-0.543	.587	-.096

Table 3
Distribution of participants by highest factor loading on F1 and F2 pre- and post-test.

		After intervention		
		F1	F2	Total
Before intervention	F1	8	0	8
	F2	6	2	8
	Total	14	2	16

tended to display beliefs aligning more strongly with profile F1 after the intervention. The mismatch with the results of the Wilcoxon signed-rank test, which revealed no significant differences in the overall strength of participants' pretest-posttest mean factor loadings on F1, can be explained in the following way: whilst the very high factor loading of some of the participants who initially loaded on F1 slightly decreased after the intervention, this did not cause any change in participants' *overall* alignment with the factor; on the other hand, participants who initially loaded weakly and non-significantly on F1 saw a considerable increase in their factor loadings, which was large enough to result in a change of factor alignment.

Method 3: Q-factor analysis of post-intervention Q-sorts and comparison of factors between time points

A new Q-factor analysis was conducted using participants' post-test Q-sorts. An assessment of the first eight factors extracted through PCA (Fig. 4) against the same criteria used at baseline (see Supplementary Material) resulted in a three-factor solution. This solution explained 65% of the variance and accounted for 15 of the 16 Q-sorts. Although the first and third factor were each defined by three-to-four participants with a strong factor loading ($r > .70$), the second factor was defined by only one participant with an extremely strong factor loading ($r = .907$). This factor was nonetheless included in the solution for its connection to the second baseline factor (see below).

To assess the degree of similarity between the factors extracted at the two time points, the factor arrays of the three post-intervention factors (henceforth Post-F1, Post-F2 and Post-F3) were correlated with the factor arrays of the two baseline factors (henceforth Pre-F1 and Pre-F2). The correlation coefficients, reported in Table 4, suggest that factor Pre-F1 remained expressed in participants' post-test Q-sorts, whereas factor Pre-F2 virtually disappeared, being only moderately correlated with factor Post-F2

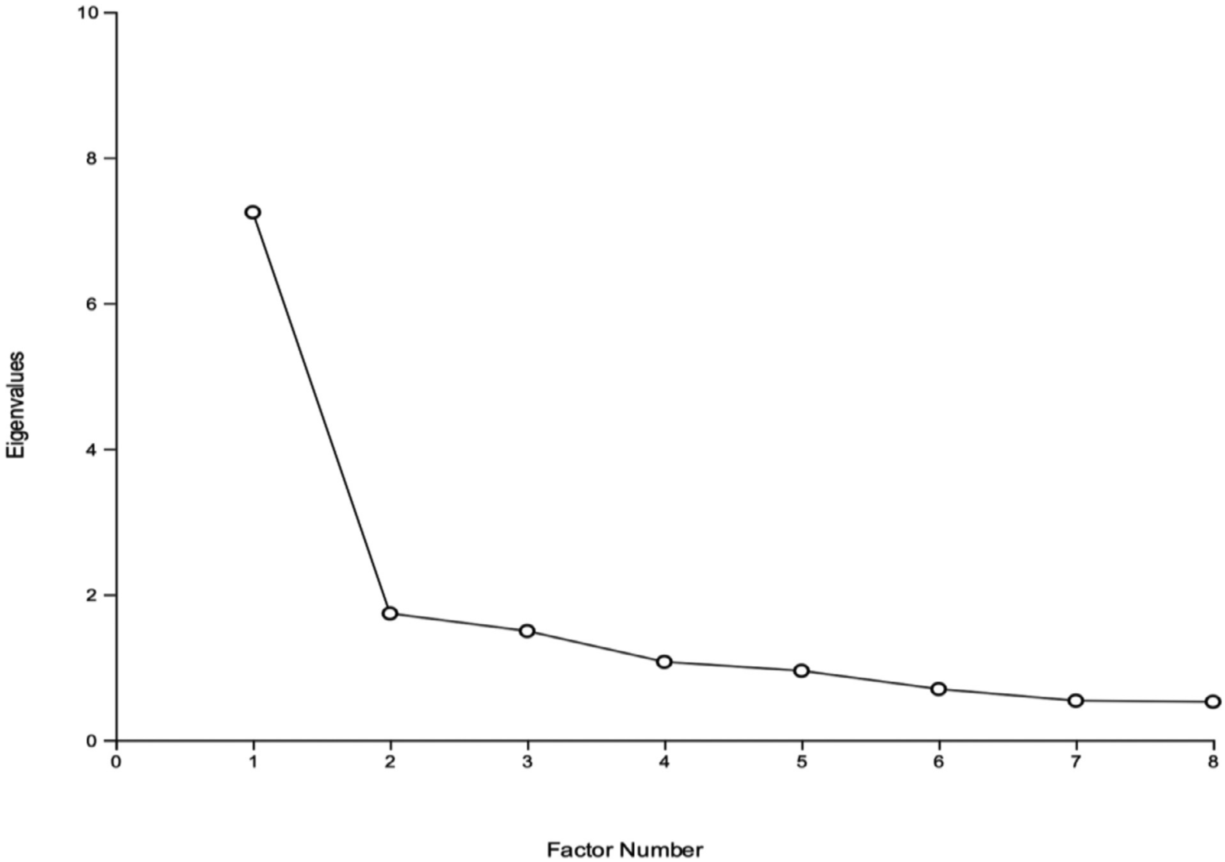

Fig. 4. Scree plot of Q-sorts (Post-Test).

Table 4
Pearson's correlation coefficients of factors pre- and post-test.

	Post-F1	Post-F2	Post-F3
Pre-F1	.786	-.073	.680
Pre-F2	.166	.493	.420

Note. $p < .05$; $p < .01$.

which, in turn, was only defined by one Q-sort. Additionally, the post-intervention factor solution revealed a new factor, Post-F3, which seems to share features of Pre-F1 and, to a lesser extent, of Pre-F2.

A comparison of the factor arrays of each pair of factors confirmed that factor Post-F1 coincides with Pre-F1, as indicated by both the high correlation coefficient and the similarities between factor scores. On the other hand, factor post-F2 bears many similarities with factor pre-F2, but it is characterised by an even stronger opposition to the use of students' other languages in school and the classroom. An analysis of the post-sorting interview with the participant helped understand the reasons behind their atypical trajectory. In summary, the participant trained in a linguistically diverse school with problems related to student behaviour, to which the school responded by policing the use of students' home languages in the classroom. The participant thus seemed to have assimilated the beliefs behind these practices, negating any potentially positive effect of the intervention.

Finally, factor Post-F3 closely aligns with factor Pre-F1, but it also shares some features of factor Pre-F2. In general, pre-service teachers aligning with this profile seem to share the view that students' other languages are a resource in school and the classroom; however, they tend to be ambivalent about allowing students to use languages other than English (or the target language) in class and about using linguistically inclusive teaching strategies. Factor Post-F3 was not represented in the pre-test factor analysis, and all the four participants who uniquely loaded on this factor used to significantly and strongly load on factor Pre-F2 before the intervention. This indicates that factor Post-F3 emerged as a result of the disappearance of factor Pre-F2. Considering the defining features of factor post-F3, it thus seems that the beliefs of almost all pre-service teachers who originally loaded on Pre-F2 became much more aligned with factors Pre-F1 and Post-F1, whilst still maintaining some elements of Pre-F2.

Method 4: Exploring individuals' shifting subjectivities

The three methods proposed above can be used to investigate change in participants' Q-sorts at the group level. As such, they enable us to detect whether the viewpoints of the group tended to change over time or after an intervention. However, the processes behind any change would still remain unclear. As a result, these approaches can be enriched by investigating individual participants' trajectories of change by analysing participants' retrospective reflections, collected through post-sorting interviews and other qualitative methods. This allows us to account for the contextual factors that may have contributed to (or hindered) any change in participants' subjectivity and understand the processes behind the trends revealed by the group-level analyses.

In the case of this exemplification, participants were interviewed after completing their second Q-sort. During the interview, participants were shown their pre- and post-intervention Q-sorts, with statements that were placed in different positions between time points highlighted. Participants were then invited to reflect on the reasons behind the changes in their Q-sorts. Additionally, during the intervention (which lasted five months), participants were asked to reflect on their experience of teaching students with diverse language backgrounds by writing a retrospective blog at the end of each school term. The retrospective data obtained from the interviews and blogs were analysed thematically and longitudinally within a multiple-case-study design, allowing to reconstruct the trajectories of belief-development experienced by selected participants. For this study, the selection of cases was guided by the trends emerged from the group-level analyses, thus focusing on participants who had moved from factor Pre-F2 to Post-F3 after the intervention. However, depending on the researcher's aims, one can select participants not aligning with the general trend: in the case of the study presented here, the only participant that remained aligned with factor Pre-F2 after the intervention would represent a clear example of an atypical trajectory.

Evaluating the proposed analytical approach

A mixed approach was used to compare pre-service teachers' Q-sorts before and after an intervention, which uncovered significant changes in participants' beliefs about linguistically inclusive teaching practices after the intervention. This exemplification illustrated how different analytical techniques can be integrated to investigate change in participants' subjectivity from different angles. On the one hand, the integration of R-based and Q-based analyses allows us to statistically test for within-group change using the initial factors as a baseline for comparison and, on the other, to track the evolution of shared configurations of beliefs over time. The results of these analyses can then be used to select participants with typical and atypical trajectories, which can be explored qualitatively in a case-study design. This allows us to complement the group-level analysis with an exploration of individuals' shifting subjectivities, in order to understand the processes behind any change found at the group level.

The above approach is not without limitations. Despite the use of inferential statistics, small sample sizes and the absence of a comparable control group would make it impossible to draw any firm conclusion from the results. Additionally, the comparison of participants' distribution on the baseline factors between time points is based on a clear-cut assignation of subjects to a single baseline factor based on their highest factor loading; as a result, the test relies on a series of approximations, as differences with regard to the strength of the factor loadings and the presence of confounded Q-sorts are not accounted for. Nonetheless, the use of different approaches may counterbalance the limitations of the single analyses.

Conclusion

Q represents a promising methodology for the field of applied linguistics as it supports a holistic approach to the study of subjectivity. When a time element is included, Q can allow an assessment of change and stability in people's perspectives. Given the paucity, across fields, of research using Q and its technique in repeated-measures designs, a systematic review of the literature was conducted to understand the analytical approaches used thus far to analyse paired Q-sorts. Although the fact that the review was carried out by a single researcher inevitably increases the possibility of selection bias, attempts were made to minimise this risk by specifying the selection criteria a priori and by examining the records selected after initial screening multiple times.

Twenty-five studies were selected for the review, the examination of which revealed a variety of analytical approaches, highlighting the lack of consensus on how Q-sorts can be analysed in experimental and observational longitudinal research. Nonetheless, it was possible to group these analytical approaches into three broad categories, characterised by the use of Q-factor analysis, descriptive statistics and inferential tests. With regard to the first category (Q-factor analysis), an approximately equal number of studies either conducted separate Q-factor analyses at each time point or performed a single Q-analysis on all Q-sorts. Whilst the latter approach may facilitate a descriptive comparison of participants' association with the extracted factor between time points, it does not provide reliable baseline data necessary to conduct inferential tests, as the factor loadings obtained are affected by the inclusion of Q-sorts from different data-collection points.

An analysis of the selected studies also revealed that a number of key analytical information and decisions were not reported or acknowledged. At the research design level, these omissions include the sampling technique used and issues around participant retention and attrition over time. Similarly, with regard to Q-factor analysis, important decisions were frequently omitted, such as the criteria used for deciding a factor solution, the number of Q-sorts selected to define a factor, and the presence and treatment of confounded Q-sorts. These decisions become even more important in studies which conducted separate Q-factor analyses at different time points, because only by using the same approach and criteria through all the phases of the Q-factor analyses can the resulting factors be meaningfully compared. Finally, when studies conducted inferential tests (e.g., ANOVA and Student's t-tests), they tended not to report whether the dataset met the assumptions required to perform such tests.

In consideration of the lack of consensus on how paired Q-sorts can be analysed, a mixed approach was proposed for analysing Q-sorts over time, which was exemplified using data from a quasi-experimental study on pre-service teachers' beliefs about multilingualism. Drawing from the findings of the systematic review, the proposed approach incorporates Q- and R-based analyses to investigate change in subjectivity from different angles. Through Q-factor analysis, it allows us to assess the evolution of factors over time by detecting the emergence of new factors and disappearance of previous ones. Through R-based inferential tests, it is possible to statistically evaluate whether participants' Q-sorts have significantly changed from a baseline. Furthermore, the findings obtained from the Q- and R-analyses can form the basis for single-case examinations of individuals' shifting subjectivities, by using participants' reflections obtained through post-sorting interviews and other retrospective qualitative methods.

This article assessed the opportunities and challenges in using Q methodology to investigate the evolution of people's perspectives over time. By reviewing the analytical procedures used in longitudinal studies adopting the Q-sort method, and by proposing analytical approaches that can be adopted in future research, it aims to provide a reference for researchers in applied linguistics who are considering using Q methodology to explore change and stability in human subjectivity.

Funding

This work was supported by a University of Cambridge Faculty of Education PhD Studentship.

Declaration of Competing Interest

I have no known conflict of interest to disclose.

Acknowledgements

I would like to thank Prof Linda Fisher for her feedback on the early version of the manuscript and for her advice during the phases of manuscript revision.

I would also like to thank the two anonymous reviewers and the editor-in-chief for their suggestions on how to improve the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rmal.2022.100025.

Appendix

Databases Accessed through the EBSCO Platform:

- British Education Index.
- Child Development & Adolescent Studies.
- Education Abstracts.
- Educational Administration Abstracts.
- Education Resource Information Centre (ERIC).
- European Views of the Americas: 1493 to 1750.
- GreenFILE.
- Library, Information Science & Technology Abstracts.
- Teacher Reference Centre.

References

Akhtar-Danesh, N. (2018). Qfactor: A command for Q-methodology analysis. *The Stata Journal, 18*(2), 432–446. 10.1177/1536867X1801800209.

Banasick, S. (2019). KADE: A desktop application for Q methodology. *The Journal of Open Source Software, 4*, 1–4. 10.21105/joss.01360.

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association, 61*(313), 228–235. 10.2307/2283057.

Bramer, W. M., de Jonge, G. B., Rethlefsen, M. L., Mast, F., & Kleijnen, J. (2018). A systematic approach to searching: An efficient and complete method to develop literature searches. *Journal of the Medical Library Association: JMLA, 106*(4), 531–541. 10.5195/jmla.2018.283.

Brown, S. R. (1980). *Political subjectivity*. Yale University Press.

Brown, S. R. (1993). A *primer* on *Q methodology*. *Operant Subjectivity, 16*, 91–138.

Caruso, M., & Fraschini, N. (2021). A Q methodology study into vision of Italian L2 university students: An Australian perspective. *The Modern Language Journal, 105*(2), 552–568. 10.1111/modl.12713.

Chaiton, M. O., Cohen, J. E., O'Loughlin, J., & Rehm, J. (2009). A systematic review of longitudinal studies on the association between depression and smoking in adolescents. *BMC Public Health, 9*(1). 10.1186/1471-2458-9-356.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). Routledge.

Expositor. (1992). A note on measuring changes in Q factor loadings. *Operant Subjectivity, 15*, 56–61.

Fraschini, N., & Park, H. (2021). Anxiety in language teachers: Exploring the variety of perceptions with Q methodology. *Foreign Language Annals, 54*(2), 341–364. 10.1111/flan.12527.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105*(2), 302–308.

Irie, K., Ryan, S., & Mercer, S. (2018). Using Q methodology to investigate pre-service EFL teachers' mindsets about teaching competences. *Studies in Second Language Learning and Teaching, 8*(3), 575–598. 10.14746/ssllt.2018.8.3.3.

Leon, A. C. (1998). Descriptive and inferential statistics. In S. Bellack, & M. Hersen (Eds.), *Comprehensive clinical psychology* (pp. 243–285). Pergamon. 10.1016/B0080-4270(73)00264-9.

Lundberg, A. (2019a). Teachers' beliefs about multilingualism: Findings from Q method research. *Current Issues in Language Planning, 20*(3), 266–283. 10.1080/14664208.2018.1495373.

Lundberg, A. (2019b). Teachers' viewpoints about an educational reform concerning multilingualism in German-speaking Switzerland. *Learning and Instruction, 64*, 1–8. 10.1016/j.learninstruc.2019.101244.

Lundberg, A., de Leeuw, R., & Aliani, R. (2020). Using Q methodology: Sorting out subjectivity in educational research. *Educational Research Review, 31*, 1–16. 10.1016/j.edurev.2020.100361.

MacIntyre, P. D., MacKay, E., Ross, J., & Abel, E. (2017). The emerging need for methods appropriate to study dynamic systems: Individual differences in motivational dynamics. In L. Ortega & Z. Han (Eds.), *Language learning & language teaching* (Vol. 48, pp. 97–122). John Benjamins Publishing Company. 10.1075/lllt.48.06mac

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. 10.1007/BF02295996.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 2021*, 372. 10.1136/bmj.n71.

Schmolck, P. (2014). *PQ Method manual*. http://schmolck.userweb.mwn.de/qmethod/pqmanual.htm.

Schroedler, T., & Fischer, N. (2020). The role of beliefs in teacher professionalisation for multilingual classroom settings. *European Journal of Applied Linguistics, 8*(1), 49–72. 10.1515/eujal-2019-0040.

Stricklin, M., & Almeida, J. (2001). *PCQ: Analysis software for Q-technique*. Retrieved from http://www.pcqsoft.com .

Watts, S., & Stenner, P. (2012). *Doing Q methodological research: Theory, method and interpretation*. SAGE Publications Ltd. 10.4135/9781446251911.

Zabala, A. (2014). Qmethod: A package to explore human perspectives using Q methodology. *The R Journal, 6*(2), 163–173. http://journal.r-project.org/archive/2014-2/zabala.pdf.

Zheng, Y., Lu, X., & Ren, W. (2019). Profiling Chinese university students' motivation to learn multiple languages. *Journal of Multilingual and Multicultural Development, 40*(7), 590–604. 10.1080/01434632.2019.1571074.

Studies included in the review

Akhtar-Danesh, N., & Wingreen, S. C. (2020). How to analyze change in perception from paired Q-sorts. *Communications in Statistics – Theory and Methods*, 1–11. 10.1080/03610926.2020.1845734.

Baron, C., Sklarwitz, S., & Coddington, N. (2021). Hidden in plain sight: Museum educators' role in teacher professional development. *Teacher Development, 25*(5), 567–584. 10.1080/13664530.2021.1897659.

Baron, C., Sklarwitz, S., Bang, H., & Shatara, H. (2019). Understanding what teachers gain from professional development at historic sites. *Theory & Research in Social Education, 47*(1), 76–107. 10.1080/00933104.2018.1489927.

Brown, S. R. (1970). Consistency and the persistence of ideology: Some experimental results. *Public Opinion Quarterly, 34*(1), 60–68. 10.1086/267773.

Brown, S. R. (1977). Political literature and the response of the reader: Experimental studies of interpretation, imagery, and criticism. *American Political Science Review, 71*(2), 567–584. 10.1017/S0003055400267476.

Cook, T. J., Scioli, F. P., Jr., & Brown, S. R. (1975). Experimental design and Q-methodology: Improving the analysis of attitude change. *Political Methodology, 1*(2), 51–70.

Cuppen, E. (2012). A quasi-experimental evaluation of learning in a stakeholder dialogue on bio-energy. *Research Policy, 41*(3), 624–637. 10.1016/j.respol.2011.12.006.

Davies, B. B., & Hodge, I. D. (2012). Shifting environmental perspectives in agriculture: Repeated Q analysis and the stability of preference structures. *Ecological Economics, 83*, 51–57. 10.1016/j.ecolecon.2012.08.013.

Dennis, K. E., & Goldberg, A. P. (1996). Weight control self-efficacy types and transitions affect weight-loss outcomes in obese women. *Addictive Behaviors, 21*(1), 103–116. 10.1016/0306-4603(95)00042-9.

Freie, J. F. (1997). The effects of campaign participation on political attitudes. *Political Behavior, 19*(2), 133–156. 10.1023/A:1024858108803.

Irie, K., & Ryan, S. (2014). Study abroad and the dynamics of change in learner L2 self-concept. In Z. Dörnyei, P. D. MacIntyre, & A. Henry (Eds.), *Motivational dynamics in language learning* (pp. 343–366). Multilingual Matters. 10.21832/9781783092574-022.

Niemeyer, S. (2004). Deliberation in the wilderness: Displacing symbolic politics. *Environmental Politics, 13*(2), 347–372. 10.1080/0964401042000209612.

Niemeyer, S., & Dryzek, J. S. (2007). The ends of deliberation: Meta-consensus and inter-subjective rationality as ideal outcomes. *Swiss Political Science Review, 13*(4), 497–526. 10.1002/j.1662-6370.2007.tb00087.x.

Niemeyer, S., Ayirtman, S., & HartzKarp, J. (2013). Understanding deliberative citizens: The application of Q methodology to deliberation on policy issues. *Operant Subjectivity, 36*(2), 114–134.

Pelletier, D., Kraak, V., McCullum, C., Uusitalo, U., & Rich, R. (1999). The shaping of collective values through deliberative democracy: An empirical study from New York's North Country. *Policy Sciences, 32*, 103–131.

Popovich, M. N., & Massé, M. H. (2005). Individual assessment of media writing student attitudes: Recasting the Mass Communication Writing Apprehension Measure. *Journalism & Mass Communication Quarterly, 82*(2), 339–355. 10.1177/107769900508200207.

Popovich, M., Massé, M., & Pitts, B. (2003). Revisiting student writer apprehension: A new interpretation of the Riffe and Stacks's Writing Apprehension Measure. *Operant Subjectivity, 26*(3), 88–111.

Ramlo, S. (2019). Examining urban, American, middle-school students' divergent views of nature before and after a field trip to a university field station and nature preserve. *The Urban Review, 51*(2), 231–246. 10.1007/s11256-018-0473-x.

Richardson, L. A., Fister, C. L., & Ramlo, S. E. (2015). Effect of an exercise and weight control curriculum: Views of obesity among exercise science students. *Advances in Physiology Education, 39*(2), 43–48. 10.1152/advan.00154.2014.

Schick, K., Gartmeier, M., & Berberat, P. O. (2021). Senior medical student attitudes towards patient communication and their development across the clinical elective year – A Q-methodology study. *Frontline Learning Research, 9*(1), 1–29.

Schwartz, A., Peacock, N., McRae, K., Seymour, R., & Gilliam, M. (2010). Defining new categories of pregnancy intention in African-American women. *Women's Health Issues, 20*(6), 371–379. 10.1016/j.whi.2010.06.005.

Tornwall, J., & McDaniel, J. (2022). Key strategies in scholarly writing instruction for doctor of nursing practice students: A Q-methodology study. *Nurse Education Today, 108*, 1–8. 10.1016/j.nedt.2021.105192.

Willoughby, M. J. (1986). Nuclear freeze attitude structures: A pre- and post-test of the impact of the Korean airliner incident. *Social Science Quarterly, 67*(3), 534–544.

Wilson, D. D. (2006). Revealing shifts in attitude among undergraduates participating in academic service learning programs. *Operant Subjectivity, 30*(1–2), 23–51.

Zheng, Y., Lu, X., & Ren, W. (2020). Tracking the Evolution of Chinese Learners' Multilingual Motivation Through a Longitudinal Q Methodology. *The Modern Language Journal, 104*(4), 781–803. 10.1111/modl.12672.