Improving Data in Electronic Surveys (IDES) Final Report

Mojca Rožman[1] Andrés Christiansen[1] Rolf Strietholt[1]
Jeppe Bundsgaard[2] Julian Fraillon[1] Ronny Scherer[3]

[1]IEA
[2]Aarhus University
[3]University of Oslo
31.1.2024

Contents

Q-sort vs. Likert scale: Evidence from an international educational teacher survey

Abstract

International large-scale assessments administer context questionnaires to students, teachers, and principals to collect information about school, classroom and learning conditions. These questionnaires usually consist of a series of rating type items which often face issues such as social desirability, self-presentation, and acquiescence bias. To address these issues, in the field trial of the International Computer and Information Literacy Study (ICILS), Q-sort was introduced as an alternative question type to assess teachers' beliefs about teaching. Q-sort is a technique that was initially developed for clinical interviews, requiring respondents to arrange a series of cards according to their preference. In this paper, we use teacher data from 28 eductional systems participating in ICILS 2023 field trial to investigate the differences between two question formats, Q-sort and rating scales both reflecting the same teaching beliefs, using multiple criteria of data quality. We compare the amount of missing data, examine the distribution of responses, item and scale means, and the correlations between the scale scores and teacher characteristics. We observed that the proportion of missing values was higher for Q-sort possibly because the cognitive load is higher for the parallel sorting of a total of 18 items than for the rating items that are answered individually. In addition, we observed more variance in the Q-sort than in the rating version. The ranking of items in Q-sort removes the possibility that respondents can agree equally with all statements and and has thus the potential to reduce acquiescence bias.

1 Introduction

1.1 Measuring Attitudes in Surveys

The study of attitudes across different disciplines has been of interest for many researchers. An attitude represents a person's overall evaluation of people, groups, ideas, and objects. Attitudes contain three components: an affective, that is represented by feelings and emotions associated with an attitude object, a behavioral, that involves the mental representation of our actions regarding an attitude object, and a cognitive, that is represented by beliefs, thoughts and attributes associated with an attitude object (Eagly & Chaiken, 1998).

Attitudes are assessed in surveys, which represent a research strategy in which quantitative information is systematically collected from a sample taken from a population (Hox, de Leeuw, & Dillman, 2008a). Information about attitudes can be collected using different instruments. For example, data can be collected through questionnaires, interviews, observations, experiments, or focus groups. But before starting the data collection and deciding on an instrument that measures attitudes, the construct in focus must be conceptualized and operationalized. This means the definition of the construct, description of its properties and scope, must be determined and important subdomains of its meaning might need to be defined. Furthermore, the process of operationalization involves choosing empirical indicators of the constructs and subdomains in focus. The indicators are then part of the research instrument. Schwarz, Knäuper, Oyserman, and Stich (2008) found that one of the components shaping respondents' answers is the research instrument.

Various factors have been found to impact a person's response to a question. Some factors are related to the respondent and some to the instrument. The action of responding is a cognitive process that can be demanding. First, the respondent must understand the question. Then the respondent has to retrieve the relevant information from his memory, and consider it to form a judgement about what the answer to the question is, and finally map this answer to the response format or options provided. When the respondent arrives at an answer, the next step is to communicate it. To this extent, the instrument of data collection and contents of the answer matter. For example, in an interview, where the interviewer records the person's response, it is likely that a respondent will give a socially desired answer relative to a questionnaire, where the recording of the response usually happens further away from the administrator. Furthermore, also the type of response options available, if any, can impact the responding. For instance, close response formats can constrain the meaning of the inferred question. Respondents usually need to rely on various estimation strategies to form their answer. Closed response formats present a preselected set of responses that are presented with the question. These response scales introduce a systematic bias into the response (Schwarz et al., 2008). It was found that the response options shape the response and that the response is highly context dependent.

Specifically, context effects were observed in measurement of attitudes.

1.2 Instruments in Large-scale Assessments

Assessing attitudes with questionnaires is standard practice in social sciences. International large-scale assessments, such as the International Computer and Information Literacy Study (ICILS) and the Trends in International Mathematics and Science Study (TIMSS), administer context questionnaires to students, parents, teachers, and other school staff to collect information about school, classroom, and learning conditions. Most international large-scale assessments rely on self-administered questionnaires. Questionnaires represent a cost-efficient way to collect large amounts of data in a relative short time frame (Schulz, 2014). In addition, the data in most cases do not involve much further processing to be ready for analyses. Behavior and attitudes tend to be measured by administering sets of items to respondents. These item sets share the question stem and response options. The response options are typically represented by rating scales with ordered response options (e.g., never/sometimes/always; strongly disagree/disagree/agree/strongly agree).

There are some examples of using alternatives to rating scales in international large-scale assessments. Kyllonen and Bertling (2013) described the use of innovative item types as alternatives to rating scales that were used in the Program for International Student Assessment (PISA) in the 2012 cycle. They argued that observed differences in attitudes between countries might reflect true differences or they might reflect differences in response style. They found that when analysing the correlation between attitudes and achievement, anchoring vignettes, forced-choice methods and signal detection type corrections all showed significant correlations. Their findings suggest that response style effects mask the true relation among constructs, and with alternatives to rating scales we can come closer to the true construct value.

Other examples of the use of alternative question formats can be found in IEA's International Civic and Citizenship Education Study (ICCS) and the Teaching and Learning International Study (TALIS) Starting Strong Survey 2018. The field trial of ICCS included forced choice formats (Schulz and Carstens, 2014), and TALIS Starting Strong included situational judgement tasks (SJT) in the main survey. SJTs intend to measure the construct in a different way. While they are also self-administered questions, they describe concrete examples in the question stem reflecting real life professional contexts and provide the respondent with several options on how to address the given situation. The response options represent alternative behavioral examples that are more or less appropriate to address the described situation.

1.3 Rating versus ranking

The literature suggests that when measuring attitudes with rating scales, respondents' answers are usually subject to lack of variation, floor or ceiling effects, or low reliability of responses (Fraillon, Ainley, Schulz, Friedman, & Duckworth, 2020; Mullis, Martin, Foy, Kelly, & Fishbein, 2020; OECD, 2019; Schulz et al., 2018). These issues are referred to as social desirability, self-presentation, and acquiescence bias, and they are well documented in survey research (Lelkes & Weiss, 2015; Schaeffer & Dykema, 2020; Yannakakis & Martínez, 2015). While the rating scales are often used to assess constructs in focus, alternative item types are used with the intention to address some issues found with rating scales. An alternative way to collect data from respondents is to ask them to rank the items. It was found that ranking reduces the response style, and it improves data quality (Alwin & Krosnick, 1985; Krosnick & Alwin, 1988). The respondents have to compare the items and are therefore more engaged in responding. The ranking of items also introduces more variation into the responses because it forces the respondent to choose the order. But ranking can be more difficult as the cognitive load increases when the item list gets longer.

Q-sort is a specific type of ranking that has received attention in the literature in previous decades. The Q-sort technique was developed for clinical interviews, requiring respondents to arrange a series of cards according to their preference. Originally, respondents arranged cards on a table, so it was possible to move cards back and forth before deciding on a final ranking. This question format is particularly suitable when all statements have a positive connotation. The ranking removes the possibility that respondents can agree equally with all statements and can thus reduce acquiescence bias. However, this item format is hardly feasible in a paper-pencil survey, where it is difficult to adjust answers once they have been set, whereas in a computer-based survey it is easy to revise an initial ranking and then gradually arrive at a final order of statements. In addition, sorting statements according to personal preferences is more engaging for participants than responding to rating scales. Participants are giving a holistic perspective on the covered topic rather than isolated ratings.

The use of two different response types also affects the consequent analysis of the data. Rating items are susceptible to response style and can produce spuriously positive correlations which can complicate the analysis of the latent structure. Alwin and Krosnick (1985) found that when measuring parental values for children, the rankings and ratings produced similar results in terms of ordering the relative importance of values but dissimilar results regarding the latent structure of the construct. In addition, the two measures showed different correlations between values and theoretically relevant predictor variables. Furthermore, ranking produces a linear dependency among the set of ranked items and therefore the use of conventional statistical techniques is not always

possible. In addition, ranking can reduce social desirability and result in more clear-cut factor structures compared to Likert-scale formats (Fluckinger, 2014).

1.4 Advantages of digital mode

In the past, mostly paper-based surveys were used in which participants were asked to respond to a series of rating type items using a response scale with mostly three or four response options. Computer-based surveys promise to improve the quality of survey data as well and solve some of the aforementioned issues of rating scales by administering items or response scales that are not possible or difficult to implement on paper. Specifically, the administration of questionnaires on the computer provides an opportunity to use functions such as sliders, drag-and-drop, or drop-down menu. Another advantage of the computer-based surveys is that the respondents easier change their response. Even if large-scale assessment studies changed the data collection mode from a paper-based to a computer-based survey, the item formats and response scales in the context surveys have remained largely unchanged. In addition, previous research on the effects of item formats is inconsistent, and the results vary depending on respondents' education and age, among other factors (Krosnick & Presser, 2010; Schwarz, Knäuper, Oyserman, & Stich, 2012). Furthermore, cross-cultural differences in response behavior are well documented and can compromise the generalizability and validity of the inferences drawn on item format effects (Fons & Tanzer, 2004; van de Vijver & He, 2014). Thus, there is a need for evidence on the effects of question types for the instruments used and populations considered in IEA studies and other international comparative studies.

1.5 The present study

1.5.1 Teaching beliefs

In this study we focus on teachers' pedagogical teaching beliefs to assess the feasibility of using Q-sort in the development of questionnaires in IEA studies. Beliefs have been studied in various ways (OECD, 2009, 2014; Tondeur, van Braak, Ertmer, & Ottenbreit-Leftwich, 2017), often focusing on two distinct and opposing beliefs. The first one is called teacher-centered approach, instructivism or direct transmission approach, and is characterized by focus on the teacher's presentation of facts and demonstration of how to carry out tasks, and students' drill and practice of content. The second is called student-centered approach or constructivism, characterized by the teacher scaffolding and facilitating students' inquiry, development of hypotheses, and presentation of their work. But several other beliefs can be identified. For example Cochrane and Paerata (2022) mention, in addition to instructivism and constructivism, behaviourism, cognitivism, socio-constructivism, and connectivism.

In theory, teachers' pedagogical teaching beliefs determine how they teach. Research on TALIS data has found a rather weak correlation between constructivist beliefs and adapted learning environments supportive of students' construction of knowledge (OECD, 2014). Yet, it is also debated if the relationship between beliefs and practices is linear, or if there is a more complex, dialectical and multi-causal relationship. However, if it is possible to achieve well-founded support for relationships between teachers' pedagogical beliefs, their practice, and students' outcome, it will be of considerable importance for teacher education and professional development, design of curriculum and teaching material among others.

Pedagogical teaching beliefs are usually assessed with ordinary Likert scale items (Safrudiannur & Rott, 2019). Several authors suggest that there is a tendency for strong social desirability in Likert scales of teaching beliefs (Di Martino & Sabena, 2010; Safrudiannur & Rott, 2019), and often a very high agreement to statements from opposing beliefs. For example, we find this using data derived from TALIS 2008 (results are available upon request). Additionally, in several of the participating countries there was even a positive correlation between the two, theoretically opposing views (OECD, 2009). Therefore, it is worth assessing the construct with an alternative question format.

1.5.2 Research questions

The ranking and rating response scales have been already compared by researchers from different fields (Coe, 2002; Harzing et al., 2009; Moors, Vriens, Gelissen, & Vermunt, 2016; Sung & Wu, 2018; van Herk & van de Velden, 2007). The differences in study settings and the possible presence of dependencies between items make it difficult to compare results across studies. In this paper, we investigate the feasibility of using the Q-sort format when collection information about teaching beliefs in an international survey and explore the quality and usefulness of the data gathered by two question types.

We examine the effect of question format using multiple criteria of data quality. First, we compare the amount of missing data for the Q-sort and rating format items. Likert scale is a specific rating scale, and we will refer in further text only to Likert scales as they are the most used rating scales. We hypothesize that the proportion of missing values is higher for Q-sort, because the cognitive load is higher for the parallel sorting of a total of 18 items than for the Likert items that are answered individually. Second, we examine the distribution across responses to the same item by different respondents. Typically, smaller variability is found in Likert items due to acquiescence, that is perceived as a problem of Likert-items. The Q-sort format is expected to reduce the tendency of acquiescence because individuals are forced to choose an order for the items. We compare Q-sort and Likert items by comparing their item and scale means. Third, we examine the correlations between the items as well as the scale scores that are constructed in a same way and transformed to be on the same scale.

2 Method

2.1 Data

For the present study, we used data from the ICILS 2023 field trial. The ICILS 2023 field trial was conducted from March to June 2022, and we used the data of 28 participating educational systems with a total sample size of 9042 secondary school teachers. The ICILS field trial questionnaire had two alternative versions, these were randomly assigned to teachers within schools with a total count of 4882 teachers for the Q-sort version and 4160 for the Likert version.

2.2 Measures

The ICILS 2023 field trial measured teachers' epistemological beliefs by presenting the teachers 18 statements (from A to R) stating their beliefs about knowledge, learning and cognition. There were six statements representing each of the three different epistemological beliefs:

- cognitivism (statements A, D, F, I, L, R),
- constructivism (statements J, K, M, N, O, Q), and
- embodied cognition (statements B, C, E, G, H, P).

Teachers answering the Q-sort questionnaire, were asked to sort 17 out of 18[1] statements into 5 available ranks (with a forced distribution among these ranks): two Rank 1 (*most closely reflects my beliefs*), four Rank 2, five Rank 3, four Rank 4, and two Rank 5 (*least closely reflects my beliefs*).

Teachers answering the Likert format were asked to mark on a six-point scale to what extent the statements them reflect their beliefs (*not at all*, *to a very small extent*, *to a small extent*, *to a moderate extent*, *to a large extent*, *completely*).

To compare both approaches we obtained numerical scores for each statement and calculated an average per epistemological belief. Moreover, to enhance these comparisons we coded both versions using a common scale with a minimum of -2 and a maximum value of 2, where 0 represents middle preference. Therefore, Q-sort original ranks were transformed as follows: Rank 1 into 2, Rank 2 into 1, Rank 3 into 0, Rank 4 into -1, and Rank 5 into -2. Alongside, the Likert answers were coded sequentially from 2 (*completely*) to -2 (*not at all*) in 5 steps of 0.8 points.

[1]Due to technical limitations only 17 ranks were implemented, leaving 1 statement behind. This was explicitly mentioned in the instructions. Teachers decided which statement they would not rank.

2.2.1 Covariates

To explain the patterns of response we used the following covariates: gender, age, time using ICT and teaching subject as these were administered in both questionnaire versions.
Time using ICT. Teachers were presented with the question *Approximately how long have you been using ICT for teaching purposes?*, the items were *During lessons, Preparing lessons, After lessons (e.g., for marking student work or reporting student learning progress)*, and the response options 1=Never;2=Less than two years;3=At least 3 but less than 5 years;4=At least 5 years but less than 10 years;5=10 years or more. We created an average score of the three items where higher values represent more time using ICT for teaching.
Teaching subject. Teachers were presented with the *What are the main subjects that you teach in this school in the current school year?*, with possible response options *[Language arts: test language], [Language arts: foreign or other national languages], Mathematics, Sciences (general science and/or physics, chemistry, biology, geology, earth sciences, technical science, etc), Human sciences / Humanities / Social studies (history, geography, civics, law, economics, etc), Creative arts (visual arts, music, dance, drama, etc), [Information technology, computer studies or similar], Practical and vocational subjects [add any appropriate national examples], Other (e.g., [moral/ethics, physical education, personal and social development])*. The respondents could choose as many responses as applicable. We focused on the first five subjects only.

2.3 Procedure

Data from both questionnaires were compared within a three step process:

1. We analyzed non-response percentages for all participants, by gender, by statement, and by epistemological belief.

2. We analyzed the distribution of the responses by comparing the mean scores by statement, and epistemological belief.

3. And finally, we analyzed the associations of responses within the scales (and between epistemological beliefs), and with teacher characteristics, using bivariate correlations and linear regressions, respectively.

3 Results

3.1 Non-response rates

The missing rates for each of the two question formats were computed and are presented in Figure 1. The Q-sort format requires special consideration regarding its missing patterns,

therefore teachers' answers were classified as: a) completed, if the teacher filled the 17 fields without repeating any letter; b) incomplete ("with missings") if the teacher used at least one statement but did not complete the grid; and c) empty, if the teacher did not classified any statement of the Q-sort section.

As Figure 1 shows, the Q-sort produces a lower amount of valid responses, that is 37% of teachers omitted the Q-sort format altogether against only 3% of teachers the Likert format. Moreover, 21% of teachers answered the Q-sort format incompletely against 7% the Likert format. Thus, using the Q-sort format we can obtain complete information only about half of the teachers in comparison with the Likert format (42% vs. 89%). Interestingly, we found different omission rates by gender in the Q-sort format. Male teachers omitted the Q-sort question more often than female teachers.

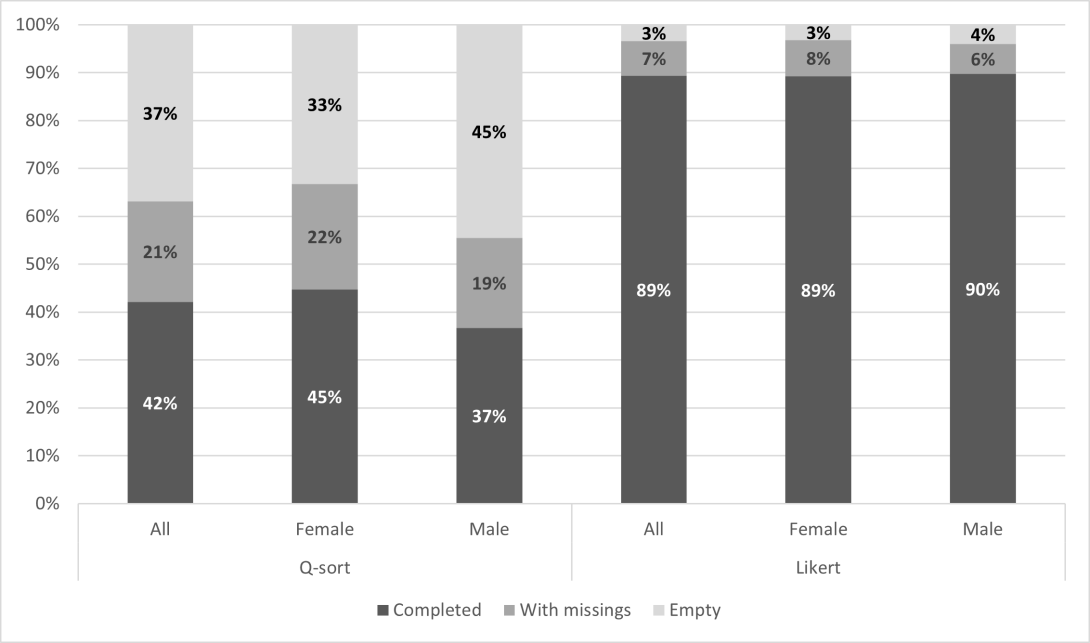

Figure 1: Percentage of missings by question type and gender

Figure 2 shows the percentage of missing items among the teachers with incomplete responses for each of the two versions. The most incomplete cases in the Likert version are missing one response, a bit less than half. For the Q-sort version we observe that half of the incomplete responses are missing two statements.

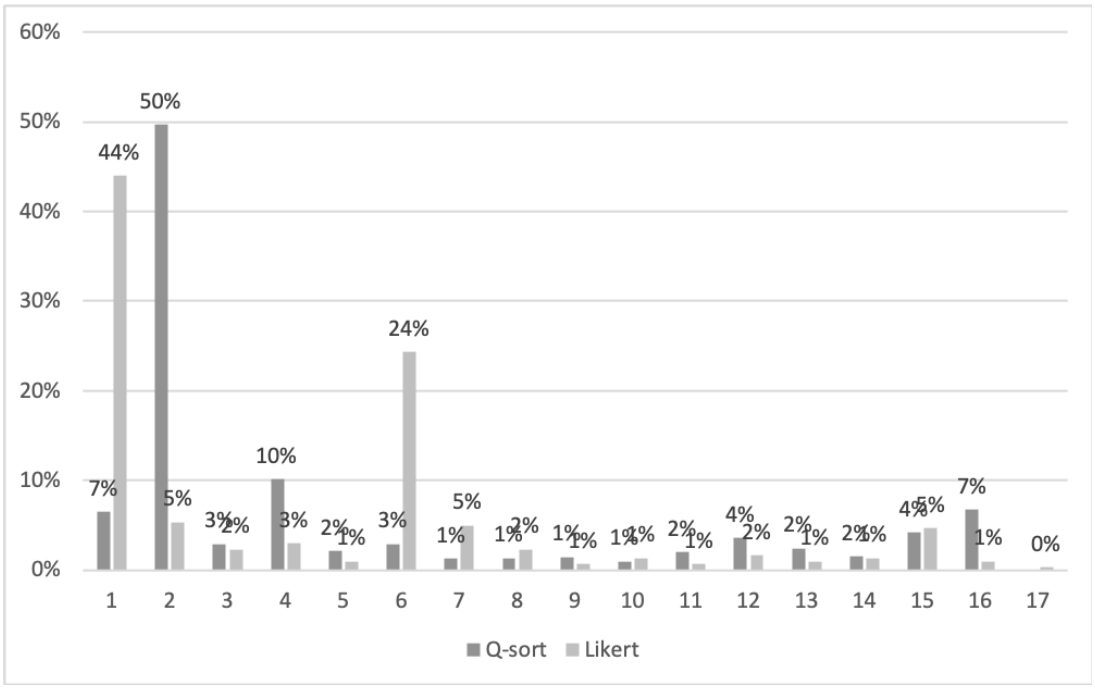

Figure 2: Number of statements missing a response by question type

In Q-sort the respondents had to omit one statement because there was one box less than the number of statements (17 boxes for 18 statements). In Table 1 we present in how many cases which statement was left out. We can observe a tendency that the respondents more often omitted the last three statements (P, Q, R; 13% to 23%). The percentage of omission for the first 15 statements is considerably lower, between 1% and 7%. For the Likert version we observe an omission rate between 1% and 3% for each statement, regardless of its position.

Table 1: Percentage of missings by epistemological belief and statement

Epistemological belief	Q-sort		Likert	Statement	Position	Q-sort		Likert
	Completed	Non completed				Completed	Non completed	
Cognitivism	6.3	32.9	2.0	A	1	2.4	22.0	1.3
				D	4	4.4	34.6	2.2
				F	6	1.7	32.4	1.7
				I	9	3.0	35.0	1.9
				L	12	3.5	31.4	1.6
				R	18	22.9	42.0	3.3
Constructivism	5.8	35.6	1.8	J	10	3.9	35.0	1.8
				K	11	3.4	34.3	1.7
				M	13	3.4	32.4	1.4
				N	14	3.8	35.6	1.7
				O	15	6.6	38.6	2.1
				Q	17	13.9	37.4	2.4
Embodied cognition	4.5	34.2	1.7	B	2	1.8	28.2	1.4
				C	3	2.2	29.1	1.5
				E	5	6.0	37.2	2.2
				G	7	2.7	33.0	1.4
				H	8	1.5	34.0	1.5
				P	16	12.7	44.1	2.3
Sample size	2059	1025	4019			2059	1025	4019

3.1.1 Non-response by teacher characteristics

Furthermore, we examined the missings in relation to other teacher characteristics. As Table 2 shows, there is a statistically significant relationship between gender and the percentage of missing responses in the Q-sort format. We also found small significant relations between the missing percentage, for the Q-sort format, and the time using ICT and the teaching subject.

Table 2: Simple regression coefficients using the proportions of missing values as criterion variable

Covariate	Q-sort	Likert
Woman	−0.108	−0.004
Age	−0.005	0.002
Time using ICT	−0.057	−0.008
Teaching subject test language	0.014	−0.009
Teaching subject foreign language	−0.079	0.006
Teaching subject maths	−0.064	0.002
Teaching subject science	−0.041	−0.011
Teaching subject informational technology	−0.034	0.010
Sample size	4882	4160

Estimates with $p \geq 0.05$ are shown in gray.

3.2 Distribution of item responses and mean scores

As stated before, we transformed the original response options of the two question version to be on the same scale with a minimum of -2 and a maximum value of 2, where 0 represents middle preference. In Table 3 we present the average values by construct and question version. For the Likert version all the average values are between 0.74 and 1.5 but for the last three sentences, where we observe the lowest average agreement. Consistently we observed the lowest preference for the last three sentences in Q-sort. But in general the average values are more spread than in the Likert question type.

Table 3: Mean scores by epistemological belief and statement

Epistemological belief	Q-sort		Likert	Statement	Position	Q-sort		Likert
	Completed	Non completed				Completed	Non completed	
Cognitivism	0.10	0.09	0.87	A	1	0.80	0.91	0.89
				D	4	−0.14	−0.11	0.72
				F	6	0.69	0.47	1.13
				I	9	−0.01	−0.06	0.86
				L	12	0.29	0.27	1.17
				R	18	−1.05	−0.95	0.43
Constructivism	−0.28	−0.18	0.76	J	10	0.04	0.11	0.84
				K	11	0.17	0.20	0.91
				M	13	−0.26	−0.19	0.83
				N	14	−0.01	0.08	0.98
				O	15	−0.23	−0.10	0.89
				Q	17	−1.37	−1.14	0.09
Embodied cognition	0.10	0.10	0.75	B	2	0.48	0.37	0.79
				C	3	0.58	0.55	0.88
				E	5	−0.50	−0.42	0.54
				G	7	0.24	0.26	0.74
				H	8	0.62	0.50	1.08
				P	16	−0.83	−0.69	0.44
Sample size	2059	1025	3719			2059	1025	3719

In addition, we observed that the average response in the Likert version is around one for all the three constructs, with cognitivism having the highest average. As expected, we saw the tendency of respondents to agreeing to statements with no big difference for the three constructs. For the Q-sort version, we observed positive average values of the same size for cognitivism and embodied cognition, and negative for constructivism.

3.3 Associations within scales, between scales, and with teacher characteristics

3.3.1 Associations within and between scales

Tables 4 and 5 are presenting the correlations between the items for both question types. For the Q-sort we expected positive correlations of items within a construct and negative between items from different constructs, if respondents have a strong preference or aversion for one family of beliefs. It can be seen from the table that the correlations for items within constructs were closer to zero than for items from different constructs where we observed more correlations with a negative sign. But in general no big differences were found in the correlation patterns. For the Likert version, we found small and medium correlations between all items with no specific pattern regarding the three different constructs.

Table 4: Pearson correlations coefficients between Q-sort items ($N = 2059$)

	Cognitivism						Constructivism						Embodied cognition					
	A	D	F	I	L	R	J	K	M	N	O	Q	B	C	E	G	H	P
A	1.00	0.20	0.06	-0.03	-0.11	-0.03	-0.15	-0.19	-0.19	-0.23	-0.20	-0.02	0.12	-0.02	0.07	-0.11	-0.16	-0.21
D	0.20	1.00	0.04	-0.08	-0.16	-0.02	-0.19	-0.19	-0.14	-0.27	-0.20	-0.15	0.12	0.07	0.20	-0.07	-0.13	-0.21
F	0.06	0.04	1.00	0.02	0.03	-0.09	-0.14	-0.16	-0.08	-0.16	-0.08	-0.10	-0.07	-0.11	-0.04	-0.05	-0.03	-0.17
I	-0.03	-0.08	0.02	1.00	0.02	-0.08	-0.01	-0.02	-0.10	-0.08	-0.04	-0.04	-0.12	-0.08	-0.13	-0.08	0.02	-0.16
L	-0.11	-0.16	0.03	0.02	1.00	-0.06	-0.08	-0.01	0.09	0.04	0.07	0.02	-0.19	-0.26	-0.22	-0.14	0.02	-0.06
R	-0.03	-0.02	-0.09	-0.08	-0.06	1.00	-0.11	-0.10	-0.01	-0.07	0.03	0.21	-0.12	-0.13	-0.15	-0.09	-0.09	-0.04
J	-0.15	-0.19	-0.14	-0.01	-0.08	-0.11	1.00	0.33	-0.07	0.00	-0.10	-0.12	-0.13	-0.12	-0.09	0.04	-0.02	-0.03
K	-0.19	-0.19	-0.16	-0.02	-0.01	-0.10	0.33	1.00	-0.02	0.05	-0.09	-0.10	-0.15	-0.08	-0.12	-0.09	-0.05	0.00
M	-0.19	-0.14	-0.08	-0.10	0.09	-0.01	-0.07	-0.02	1.00	0.06	0.10	0.02	-0.17	-0.13	-0.19	-0.09	-0.01	0.00
N	-0.23	-0.27	-0.16	-0.08	0.04	-0.07	0.00	0.05	0.06	1.00	0.12	0.01	-0.20	-0.11	-0.16	-0.05	-0.06	0.17
O	-0.20	-0.20	-0.08	-0.04	0.07	0.03	-0.10	-0.09	0.10	0.12	1.00	0.08	-0.16	-0.16	-0.21	-0.07	-0.03	0.13
Q	-0.02	-0.15	-0.10	-0.04	0.02	0.21	-0.12	-0.10	0.02	0.01	0.08	1.00	-0.16	-0.13	-0.18	-0.12	-0.08	0.07
B	0.12	0.12	-0.07	-0.12	-0.19	-0.12	-0.13	-0.15	-0.17	-0.20	-0.16	-0.16	1.00	0.15	0.21	-0.07	-0.15	-0.10
C	-0.02	0.07	-0.11	-0.08	-0.26	-0.13	-0.12	-0.08	-0.13	-0.11	-0.16	-0.13	0.15	1.00	0.12	-0.02	-0.05	-0.03
E	0.07	0.20	-0.04	-0.13	-0.22	-0.15	-0.09	-0.12	-0.19	-0.16	-0.21	-0.18	0.21	0.12	1.00	-0.05	-0.10	-0.12
G	-0.11	-0.07	-0.05	-0.08	-0.14	-0.09	0.04	-0.09	-0.09	-0.05	-0.07	-0.12	-0.07	-0.02	-0.05	1.00	-0.01	-0.06
H	-0.16	-0.13	-0.03	0.02	0.02	-0.09	-0.02	-0.05	-0.01	-0.06	-0.03	-0.08	-0.15	-0.05	-0.10	-0.01	1.00	-0.10
P	-0.21	-0.21	-0.17	-0.16	-0.06	-0.04	-0.03	0.00	0.00	0.17	0.13	0.07	-0.10	-0.03	-0.12	-0.06	-0.10	1.00

Estimates with $p \geq 0.05$ are shown in gray.

Table 5: Pearson correlations coefficients between Likert items ($N = 3719$)

	Cognitivism						Constructivism						Embodied cognition					
	A	D	F	I	L	R	J	K	M	N	O	Q	B	C	E	G	H	P
A	1.00	0.56	0.60	0.39	0.38	0.24	0.32	0.31	0.36	0.37	0.35	0.10	0.58	0.53	0.41	0.37	0.45	0.16
D	0.56	1.00	0.57	0.39	0.34	0.35	0.34	0.35	0.33	0.37	0.37	0.17	0.56	0.61	0.56	0.36	0.41	0.23
F	0.60	0.57	1.00	0.39	0.43	0.27	0.32	0.33	0.36	0.37	0.36	0.08	0.54	0.53	0.46	0.35	0.49	0.15
I	0.39	0.39	0.39	1.00	0.43	0.33	0.51	0.48	0.35	0.45	0.47	0.26	0.37	0.36	0.33	0.47	0.51	0.31
L	0.38	0.34	0.43	0.43	1.00	0.28	0.42	0.42	0.33	0.40	0.41	0.19	0.32	0.32	0.29	0.36	0.47	0.20
R	0.24	0.35	0.27	0.33	0.28	1.00	0.33	0.32	0.32	0.38	0.41	0.56	0.24	0.26	0.33	0.32	0.29	0.43
J	0.32	0.34	0.32	0.51	0.42	0.33	1.00	0.72	0.34	0.42	0.43	0.26	0.33	0.36	0.35	0.42	0.47	0.32
K	0.31	0.35	0.33	0.48	0.42	0.32	0.72	1.00	0.32	0.44	0.42	0.23	0.31	0.35	0.35	0.40	0.42	0.31
M	0.36	0.33	0.36	0.35	0.33	0.32	0.34	0.32	1.00	0.53	0.46	0.24	0.34	0.33	0.28	0.36	0.45	0.32
N	0.37	0.37	0.37	0.45	0.40	0.38	0.42	0.44	0.53	1.00	0.58	0.28	0.39	0.41	0.36	0.39	0.46	0.45
O	0.35	0.37	0.36	0.47	0.41	0.41	0.43	0.42	0.46	0.58	1.00	0.33	0.34	0.34	0.33	0.40	0.45	0.45
Q	0.10	0.17	0.08	0.26	0.19	0.56	0.26	0.23	0.24	0.28	0.33	1.00	0.12	0.13	0.24	0.21	0.18	0.49
B	0.58	0.56	0.54	0.37	0.32	0.24	0.33	0.31	0.34	0.39	0.34	0.12	1.00	0.58	0.51	0.34	0.45	0.24
C	0.53	0.61	0.53	0.36	0.32	0.26	0.36	0.35	0.33	0.41	0.34	0.13	0.58	1.00	0.48	0.36	0.46	0.26
E	0.41	0.56	0.46	0.33	0.29	0.33	0.35	0.35	0.28	0.36	0.33	0.24	0.51	0.48	1.00	0.34	0.35	0.29
G	0.37	0.36	0.35	0.47	0.36	0.32	0.42	0.40	0.36	0.39	0.40	0.21	0.34	0.36	0.34	1.00	0.53	0.30
H	0.45	0.41	0.49	0.51	0.47	0.29	0.47	0.42	0.45	0.46	0.45	0.18	0.45	0.46	0.35	0.53	1.00	0.29
P	0.16	0.23	0.15	0.31	0.20	0.43	0.32	0.31	0.32	0.45	0.45	0.49	0.24	0.26	0.29	0.30	0.29	1.00

Estimates with $p \geq 0.05$ are shown in gray.

Furthermore, we explored the relations between the scale scores with teacher characteristics and between the constructs. The results are presented in Table 6. The correlations between the constructs in Q-sort were negative and small in size. For the Likert version, we observed high positive correlation.

Table 6: Correlations between constructs

	Q-sort			Likert		
	A	B	C	A	B	C
A. Cognitivism	1.00	−0.35	−0.24	1.00	0.69	0.78
B. Constructivism	−0.35	1.00	−0.34	0.69	1.00	0.70
C. Embodied cognition	−0.24	−0.34	1.00	0.78	0.70	1.00
Sample size		2059			3719	

3.3.2 Associations with teacher characteristics

The regression coefficients representing the association of teaching beliefs and other teacher characteristics are presented in Table 7. The results between the two question versions are showing different patterns. For example we found a difference between male and female teachers for cognitivism in both versions. But in the Q-sort cognitivism is preferred by male and in the Likert version by female teachers. The opposite significant relationship is also found for time using ICT. Further, significant differences are found for some relations between teacher characteristics and constructivism but they point to the same direction for both question types. Finally, we again find significant coefficients pointing to different directions for embodied cognition and time using ICT. Whereas in the Q-sort

version embodied cognition is preferred by teachers that use ICT less time, in the Likert version embodied cognition is preferred by teachers that use ICT more time. In these situations our conclusions about the relations between beliefs and teacher characteristics would depend on the question format.

Table 7: Linear regression coefficients using epistemological beliefs as criterion variables

Covariate	Q-sort			Likert		
	Cog	Con	Emb cog	Cog	Con	Emb cog
Woman	−0.132	0.044	0.019	0.153	0.047	0.209
Age	−0.028	0.046	−0.042	0.095	0.060	0.032
Time using ICT	−0.066	0.052	−0.053	0.159	0.116	0.138
Teaching subject test language	−0.022	0.114	0.007	0.008	0.065	0.029
Teaching subject foreign language	−0.038	0.158	−0.070	0.061	0.026	0.047
Teaching subject maths	0.124	−0.074	−0.090	−0.063	−0.165	−0.094
Teaching subject science	0.098	−0.071	−0.044	0.035	−0.049	−0.029
Sample size		2029			3719	

Estimates with $p \geq 0.05$ are shown in gray.

4 Discussion

The primary objective of this study was to compare two variations of an instrument developed to evaluate three key teacher epistemological teaching beliefs: instructivism, constructivism, and embodied cognition. Both Likert-item and Q-sort versions of this instrument were assessed against various criteria such as missing values, distribution of responses, and correlation with additional variables to explore the nomological networks.

Our study made several critical discoveries. Firstly, non-response rates were notably higher with the Q-sort instrument, with 40% of participants not responding to any of the Q-sort items and an additional 20% submitting incomplete data. This contrasts with the Likert-based instrument, which had a less than 10% non-response rate overall. Secondly, differences were observed in the response distribution across the two instrument versions. Notably, the Likert-items version exhibited acquiescence bias, with teachers showing a tendency to agree with all three of the presented epistemological teaching beliefs. However, the Q-sort version displayed a higher degree of variance and less acquiescence bias. Finally, differing patterns of correlations emerged between the three teaching beliefs and other variables, underscoring the disparity between the two instrument versions.

The strengths of the study are manifold. The use of an experimental design boosted the solidity of the findings. Furthermore, a sizable sample of over 9000 teachers contributed to the study's statistical power and the potential for generalizing its findings. The study's utilization of international large-scale data also adds to the strength of the study's generalizability. Importantly, by investigating epistemological teaching beliefs, the study focuses on a significant variable within the realm of teacher research.

Nonetheless, the study does have its limitations. Primarily, it concentrated exclusively on epistemological beliefs, so any attempt to generalize the findings to other critical constructs should be done with caution. The placement of the items on epistemological teaching beliefs at the end of the instrument might have contributed to the higher non-response rate for this version. The study also faced technical constraints from the online survey system that limited teachers to ranking only 17 of the 18 statements, potentially affecting the Q-sort results. Lastly, the study had limited access to external variables to examine nomological networks.

In conclusion, our study enriches understanding of item formats by contrasting two versions of an instrument designed to measure them. It underscores key differences in non-response rates, response distributions, and correlations with other variables between the Likert-based and Q-sort versions. While the study has notable strengths such as its experimental design and large sample size, certain limitations, including the exclusive focus on epistemological beliefs, item positioning, and a scarcity of external variables, must be acknowledged. It is incumbent on future research to address these limitations and consider broadening the scope to include other important constructs in education.

Response scale format in student survey: drop-down menu vs. Likert

1 Introduction

A new response format was introduced to ICILS 2023 student questionnaire. This is the drop down menu, a response format that cannot be used in a paper version of the questionnaire. A drop-down is a "toggleable" menu that allows the respondent to choose one value from a predefined list. An advantage of the drop-down is that it conserves screen space, which can be used to verbally describe each answer choice. In paper versions, often only the poles of the response scale are labeled. Drop-down menus are especially useful when it is desired to get information about the same item on more than one characteristic. In this way, the statements are presented once, and responses can be entered via drop-down menus for multiple characteristics. Although they require more clicks for the respondents to submit their choice, drop-downs can present a more condensed and efficient presentation of the content. Naming the individual response alternatives should make respondents' answers more credible and valid, since a substantive meaningful label is given for each response option. If naming the response options between the poles causes them to be selected more often, undesirable distributions such as ceiling effects could be reduced.

1.1 Research questions

We aim to investigate the effect that the response scale has on data quality. We observed the amount of missing data for both response scale versions, distribution of responses and relationships with other relevant constructs.

2 Method

We use data from the ICILS 2018 and 2023 field trial to investigate the effect of the response format. We use data from countries that were common in both cycles: Denmark, France, Finland, Germany, Italy, Kazakhstan, Korea, Portugal, and Uruguay. The use of

the previous cycle data was necessary because the newly introduced response format does not have a parallel version that would use the Likert response format. In the analysis we use data from 11828 students, 4598 from ICILS 2018 Field Trial and 7230 from ICILS 2023 Field Trial.

We observe the differences in a questions that considered the frequency of ICT use at school. The question in 2018 was as follows: "*At school, how often do you use ICT during lessons in the following subjects or areas?*". The wording of the question in 2023 was: "*How often do you use ICT for your learning in your lessons at school?*". The question in 2023 had a second part that was asking about the perceived learning improvement related to ICT use.

The available response options were the same in both cycles: *I don't study this subject/these subjects, Never, In some lessons, In most lessons, I every or almost every lesson.*

In both cycles the following items were part of the question:

- Language arts: test language,
- Language arts: foreign or other national languages,
- Mathematics,
- Sciences (general science and/or physics, chemistry, biology, geology, earth sciences,
- Human sciences/humanities/social studies (history, geography, civics, law, economics, etc.),
- Creative arts ([visual arts], music, dance, drama, etc.),
- Information technology, computer studies pr similar,
- Practical or vocational,
- Other (e.g. [moral/ethics, physical education, personal and social development]).

We used the responses to following questions as student characteristics, with the response options in brackets:

- gender (0 = boy, 1 = girl)
- What language do you speak at home most of the time? (0 = Other language, 1 = Language of test)
- What is the highest level of education you expect to complete?, (1= ISCED level 6, 7 or 8, 2 = ISCED level 4 or 5, 3 = ISCED level 3, 4 = ISCED level 2, 5 = I do not expect to complete ISCED level 2)

- About how many books are there in your home? (1 = None or very few (0–10 books), 2 = Enough to fill one shelf (11–25 books), 3 = Enough to fill one bookcase (26–100 books), 4 = Enough to fill two bookcases (101–200 books), 5 = Enough to fill three or more bookcases (more than 200 books))
- How many of the following ICT are currently used in your home? Desktop computers (1 = None, 2 = One, 3 = Two, 4 = Three or more)
- ICT self-efficacy scale (higher values represent higher self-efficacy)

The data were analyzed using descriptive statistics, and regression analysis. For each covariate we estimated the a simple regression model using the percentage of missings or ICT use at school scale score as a criterion variable. In order to provide a clear interpretation, we present these results in the original metric of the percentage of missings or ICT use scale score. Moreover, non-dichotomous covariates were standardized.

3 Results

3.1 Non-response

We observed the non-response of the pooled dataset in both cycles. The results are presented in Figure 3. We distinguish between *completed* (students responded to all of the items), *with missings* (students responded to at least one item but not all), and *empty* (students did not attempt to respond to the question) responses. We can see that the non-response is higher in ICILS 2018 (Likert) compared to 2023 (Dropdown).

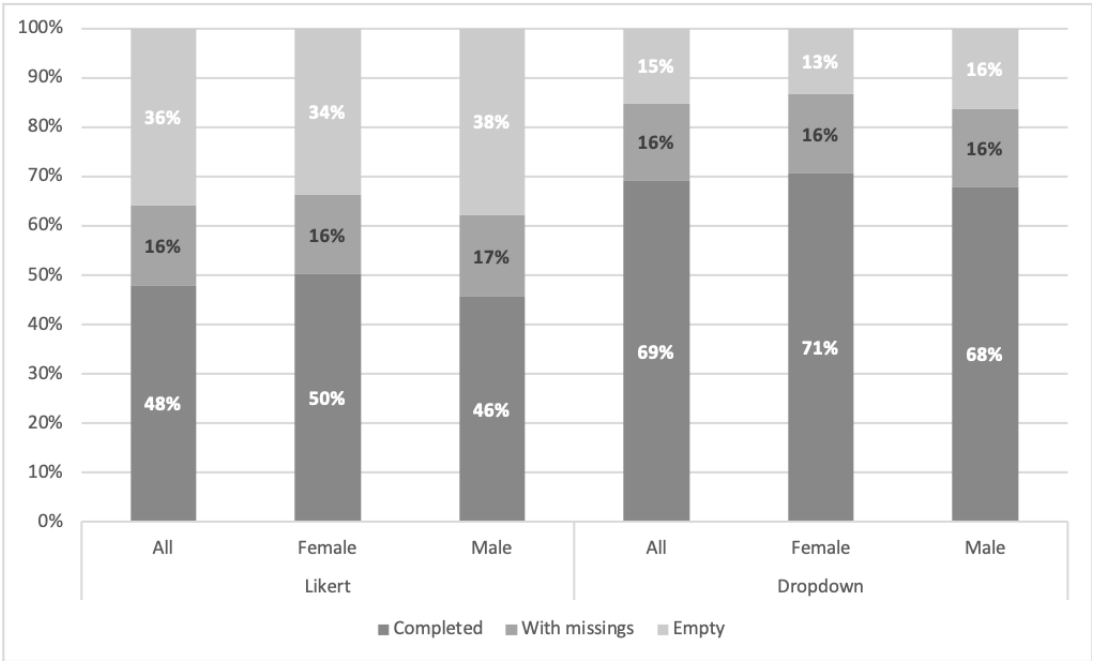

Figure 3: Percentage of missings by response format and gender

In Figure 4 we present the number of omitted items by response format. We can observe, that the vast majority of respondents in the 2023 cycle omitted only one item. In the 2018 cycle the majority omitted up to four of the nine items.

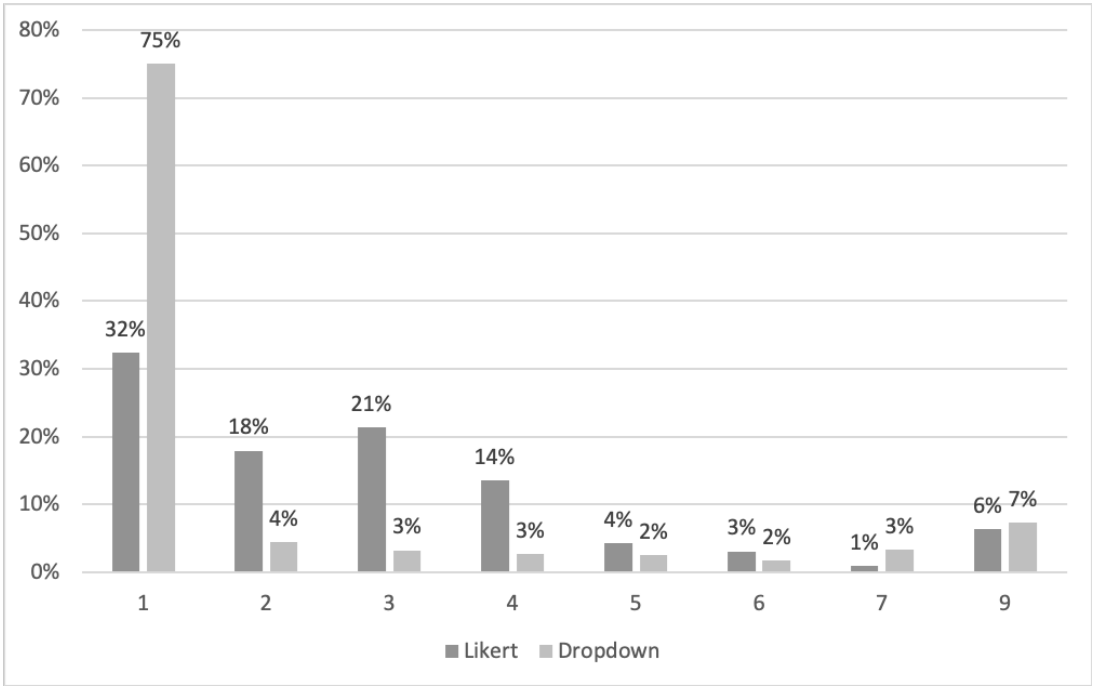

Figure 4: Number of omitted items by response format

In Table 8 we present the percentage of students that omitted each item. We considered only students that gave at least on valid response to the items. In general, the omissions are lower for the "Dropdown" compared to the "Likert" version with the exception of *practical or vocational subjects* where the omissions for the "Dropdown" are higher.

Table 8: Percentage of missings by subject

Subject	Position	Likert	Dropdown
Language arts test language	1	1.2	0.7
Language arts foreign languages	2	8.7	2.1
Mathematics	3	3.7	3.1
Sciences	4	11.7	2.9
Human sciences	5	3.8	2.9
Creative arts	6	10.4	3.3
Information technology, computer studies	7	7.6	3.6
Practical or vocational subjects	8	10.3	16.2
Other	9	14.8	4.0
Sample size		2948	6131

3.1.1 Non-response by student characteristics

Table 9: Simple regression coefficients using percentange of missings as criterion variable

	Likert	Dropdown
Girl	−0.042	−0.032
Test language	−0.052	−0.095
Expected ISCED	−0.012	−0.034
Books at home	−0.019	−0.028
Laptops at home	0.003	−0.034
ICT Self-efficacy	−0.049	−0.009
Sample size	4598	7230

Estimates with $p \geq 0.05$ are shown in gray.

3.2 Distribution

Table 10: Mean score by subject

Subject	Position	Likert	Dropdown
Language arts test language	1	0.01	−0.04
Language arts foreign languages	2	−0.02	0.02
Mathematics	3	−0.18	−0.27
Sciences	4	0.02	−0.12
Human sciences	5	−0.15	−0.06
Creative arts	6	−0.49	−0.38
Information technology, computer studies	7	−0.06	−0.07
Practical or vocational subjects	8	−0.99	−0.73
Other	9	−0.47	−0.48
Sample size		2948	6131

3.3 Associations

3.3.1 Associations betwen ICT use items

Table 11: Pearson correlations coefficients between Likert items ($N = 2948$, $\bar{\rho} = 0.331$)

	A	B	C	D	E	F	G	H	I
A. Language arts test language	1.00	0.63	0.58	0.60	0.57	0.18	0.02	0.25	0.46
B. Language arts foreign languages	0.63	1.00	0.51	0.61	0.56	0.22	0.05	0.25	0.42
C. Mathematics	0.58	0.51	1.00	0.57	0.45	0.14	0.04	0.19	0.37
D. Sciences	0.60	0.61	0.57	1.00	0.61	0.22	0.10	0.25	0.42
E. Human sciences	0.57	0.56	0.45	0.61	1.00	0.18	0.04	0.23	0.42
F. Creative arts	0.18	0.22	0.14	0.22	0.18	1.00	0.27	0.42	0.30
G. Information technology, computer studies	0.02	0.05	0.04	0.10	0.04	0.27	1.00	0.28	0.08
H. Practical or vocational subjects	0.25	0.25	0.19	0.25	0.23	0.42	0.28	1.00	0.42
I. Other	0.46	0.42	0.37	0.42	0.42	0.30	0.08	0.42	1.00

Estimates with $p \geq 0.05$ are shown in gray.

Table 12: Pearson correlations coefficients between Dropdown items ($N = 6131$, $\bar{\rho} = 0.245$)

	A	B	C	D	E	F	G	H	I
A. Language arts test language	1.00	0.52	0.48	0.44	0.46	0.12	0.02	0.18	0.30
B. Language arts foreign languages	0.52	1.00	0.43	0.44	0.44	0.13	0.04	0.15	0.28
C. Mathematics	0.48	0.43	1.00	0.54	0.35	0.08	0.02	0.14	0.24
D. Sciences	0.44	0.44	0.54	1.00	0.44	0.11	0.08	0.06	0.22
E. Human sciences	0.46	0.44	0.35	0.44	1.00	0.20	0.06	0.18	0.29
F. Creative arts	0.12	0.13	0.08	0.11	0.20	1.00	0.27	0.26	0.25
G. Information technology, computer studies	0.02	0.04	0.02	0.08	0.06	0.27	1.00	0.11	0.05
H. Practical or vocational subjects	0.18	0.15	0.14	0.06	0.18	0.26	0.11	1.00	0.40
I. Other	0.30	0.28	0.24	0.22	0.29	0.25	0.05	0.40	1.00

Estimates with $p \geq 0.05$ are shown in gray.

3.3.2 Associations with student characteristics

Table 13: Simple regression coefficients using ICT use as criterion variable

	Likert	Dropdown
Girl	0.036	0.027
Test language	−0.024	0.005
Expected ISCED	0.020	−0.006
Books at home	0.085	0.032
Laptops at home	0.031	0.065
ICT Self-efficacy	0.028	0.082
Sample size	2948	6131

Estimates with $p \geq 0.05$ are shown in gray.

4 Discussion and a note about the completion of this study

Our aim to was to investigate the effect of the response scale on data quality. Unfortunately the same question in different formats was not administered in ICILS 2023 FT. For this reason we compared the question versions from different ICILS cycles. Initial analyses showed that the amount of omissions in ICILS 2018 for the question about how often the ICT was used during lessons was unexpectedly high. If we were to compare these results with ICILS 2023 when the question was used in a different format, we could conclude that omissions are lower in 2023. Due to technical issues and limitations on the data comparability we decided to stop the analyses for this study.

Further we tried a comparison with the 2018 main survey (MS) data. We decided not to use these comparisons as there are significant differences between FT and MS that could have impacted the results.

The presented results from ICILS 2018 FT and 2023 FT could indicate that the item format contributed to less omissions as for the "Dropdown" we found less missing values than for the "Likert" version. Nevertheless, there are several issues and limitations, that we need to mention several and need to be taken into account with this interpretation:

- Item positioning effects were present. Careful inspection of missing patterns revealed item position effects but there was insufficient data to model such effects.
- Item wording was changed considerably. For 2018 items asked for use of ICT for school-related purposes in general, and for 2023, items were separated into use in and outside school.
- The different versions were tested within a different time. Unfortunately there were no parallel versions in 2023 so a comparison to 2018 was the only available option.

These issues would have had several implications on the comparability between 2018 and 2023 results. Thus, we refrained from pursuing this research project and interpretation of other results. Nevertheless, we decided to present some exploratory analyses with the available data and we caution the readers to keep in mind the aforementioned issues.

References

Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly*, *49*(4), 535–552.

Cochrane, F., & Paerata, T. (2022). 6.2 engagement and 'presence'strategies. *Teaching with Technology*.

Coe, R. (2002). Analyzing ranking and rating data from participatory on-farm trials. *Quantitative analysis of data from participatory methods in plant breeding*, 44–65.

Di Martino, P., & Sabena, C. (2010). Teachers' beliefs: The problem of inconsistency with practice. In *Proceedings of the 34th conference of the international group for the psychology of mathematics education* (Vol. 2, pp. 313–320).

Eagly, A., & Chaiken, S. (1998). Attitude structure. *Handbook of social psychology*, *1*, 269–322.

Fluckinger, C. D. (2014). Big five measurement via q-sort: An alternative method for constraining socially desirable responding. *SAGE Open*, *4*(3), 2158244014547196.

Fons, v., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). *Preparing for life in a digital world: Iea international computer and information literacy study 2018 international report*. Springer Nature.

Harzing, A.-W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., ... Zander, L. (2009). Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International business review*, *18*(4), 417–432.

Hox, J. J., de Leeuw, E. D., & Dillman (Eds.). (2008b). *International handbook of survey methodology*. Routledge.

Hox, J. J., de Leeuw, E. D., & Dillman, D. A. (2008a). The cornerstones of survey research. In J. J. Hox, E. D. de Leeuw, & Dillman (Eds.), *International handbook of survey methodology* (pp. 1–17). Routledge.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*(4), 526–538.

Krosnick, J. A., & Presser, S. (2010). Questionnaire design in: Jd wright and pv marsden. *Handbook of survey research*, 263–313.

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, 277–285.

Lelkes, Y., & Weiss, R. (2015). Much ado about acquiescence: The relative validity and reliability of construct-specific and agree–disagree questions. *Research & Politics*, *2*(3), 2053168015604173.

Moors, G., Vriens, I., Gelissen, J. P., & Vermunt, J. K. (2016). Two of a kind. similarities between ranking and rating data in measuring values. In *Survey research methods* (Vol. 10, pp. 15–33).

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science.* TIMSS & PIRLS International

Study Center, Boston College.

OECD. (2009). First results from TALIS. *Teaching and Learning International Survey*.

OECD. (2014). *TALIS 2013 results: An international perspective on teaching and learning*. Organisation for Economic Co-operation and Development.

OECD. (2019). *PISA 2018 results (volume I): What students know and can do*. Organisation for Economic Co-operation and Development.

Safrudiannur, & Rott, B. (2019). Measuring teachers' beliefs: A comparison of three different approaches. *EURASIA Journal of Mathematics, Science and Technology Education*, *16*(1), em1796.

Schaeffer, N. C., & Dykema, J. (2020). Advances in the science of asking questions. *Annual Review of Sociology*, *46*, 37–60.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2018). *ICCS 2016 international report. becoming citizens in a changing world*. Cham, Switzerland: Springer.

Schwarz, N., Knäuper, B., Oyserman, D., & Stich, C. (2008). The psychology of asking questions. In J. J. Hox, E. D. de Leeuw, & Dillman (Eds.), *International handbook of survey methodology* (pp. 18–34). Routledge.

Schwarz, N., Knäuper, B., Oyserman, D., & Stich, C. (2012). The psychology of asking questions. *International handbook of survey methodology*, 18–34.

Sung, Y.-T., & Wu, J.-S. (2018). The visual analogue scale for rating, ranking and paired-comparison (VAS-RRP): a new technique for psychological measurement. *Behavior research methods*, *50*, 1694–1715.

Tondeur, J., van Braak, J., Ertmer, P. A., & Ottenbreit-Leftwich, A. (2017). Understanding the relationship between teachers' pedagogical beliefs and technology use in education: a systematic review of qualitative evidence. *Educational technology research and development*, *65*, 555–575.

van de Vijver, F. J., & He, J. (2014). *Report on social desirability, midpoint and extreme responding in TALIS 2013*. Organisation for Economic Co-operation and Development.

van Herk, H., & van de Velden, M. (2007). Insight into the relative merits of rating and ranking in a cross-national context using three-way correspondence analysis. *Food Quality and Preference*, *18*(8), 1096–1105.

Yannakakis, G. N., & Martínez, H. P. (2015). Ratings are overrated! *Frontiers in ICT*, *2*, 13.