2-26-2019

Audio Recordings in Face-to-Face Interviews as a Means to Detect Undesirable Interviewer Behavior

Birgit Jesske
infas Institute for Applied Social Sciences, B.Jesske@infas.de

infas

How to detect undesirable interviewer behavior (UIB) during CAPI field

Labor Market and Social Security Panel (PASS)

by Birgit Jesske, infas Institute for Applied Social Sciences

RESEARCH QUESTION

Undesirable interviewer behavior (UIB) could be one source for data errors and measurement effects in the setting of standardized interviewing techniques. Survey organizations have to ensure that errors and effects are minimized by monitoring and validating their data collection processes during the entire survey period. Particularly errors caused by interviewers must be identified as early as possible. Monitoring is one method to detect undesirable interviewer behavior, which has been well established for telephone surveys from their very beginning. Monitoring face-to-face interviews is possible with listening to audio recordings, which can be easily produced in the CAPI field. How can we handle audio files in large scale surveys? How can we use them to establish a monitoring process for the CAPI field? Does the procedure detect undesirable interviewer behavior effectively?

BACKGROUND

The Labor Market and Social Security Panel (PASS) is a central data set for labor market and poverty research in Germany. The panel was established in 2006 at the Institute for Employment Research (IAB) and the annual survey waves that have been taking place since 2007 (including the addition of annual refresher samples) has now reached a size of, on average, 10,000 households with about 16,000 individuals per year. The PASS study design involves a mix of methods allowing for telephone interviews (CATI) as well as face-to-face interviews (CAPI). The CATI field employs approx. 150 interviewers per wave, the CAPI field approx. 350.
The study design of PASS includes various measures to avoid or minimize as well as monitor sources of error and effects at different levels. For CATI and CAPI a uniform standardized instrument (household questionnaire and individual questionnaire) is used. All interviewers ared jointly prepared by means of one-day training session. From Wave 10 onwards, audio recordings will also be created in the CAPI field for detecting undesirable interviewer behavior. For the present report, the implementation of the strategies can be reported by using the data from Wave 12 in PASS as an example. The surveys for Wave 12 took place between February and September 2018.

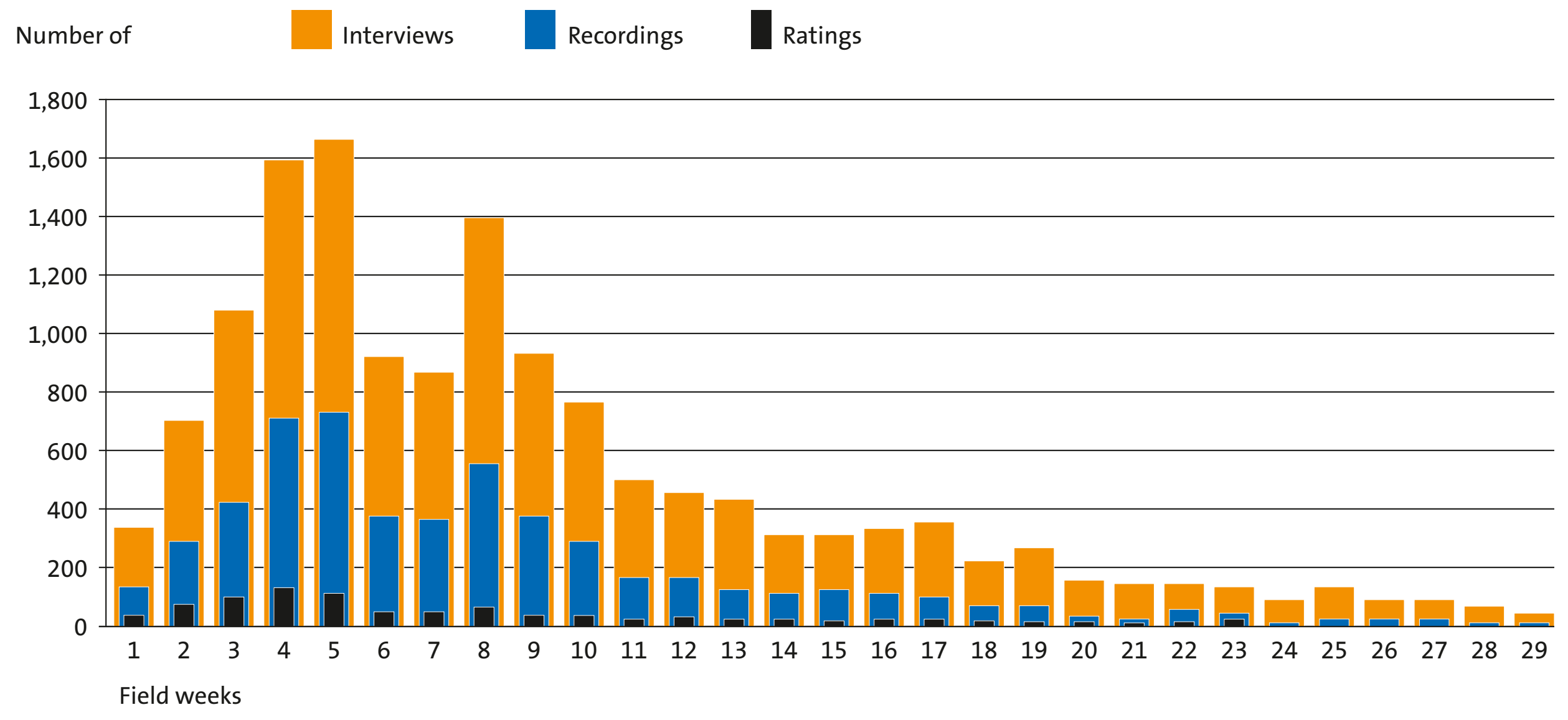
Lebensqualität
und soziale Sicherung

AUDIO RECORDINGS – OBSERVED UIB

Each interviewer should record a minimum of three interviews at the beginning and at least ten percent of all interviews. The recording of interviews in Germany requires the consent of the respondent.

Using and rating the audio recordings posed a new challenge for the survey team at infas. With the aim of ensuring standardization in the interview, deviations had to be identified as quickly as possible on the basis of the audio recordings. Interviewers with undesirable behavior should then receive feedback and be trained prior to further deployment. 794 interviews out of 5,415 audio recordings were listened to in its entire length and rated using a coding scheme (behavior coding). The coding scheme counts the number of deviations during an interview as well as documents specific examples with question numbers. The counting of deviations was unsystematic only and was not recorded for individual questions. The specific examples should be particularly useful for follow-up training and feedback discussions in order to explain the deviation to the interviewers in more detail.

FIGURE 1 Distribution of interviews, recordings and ratings over the course of the field period/weeks



14,423 interviews 5,415 recordings 749 (full) ratings

Number of interviews, recordings and full ratings over all

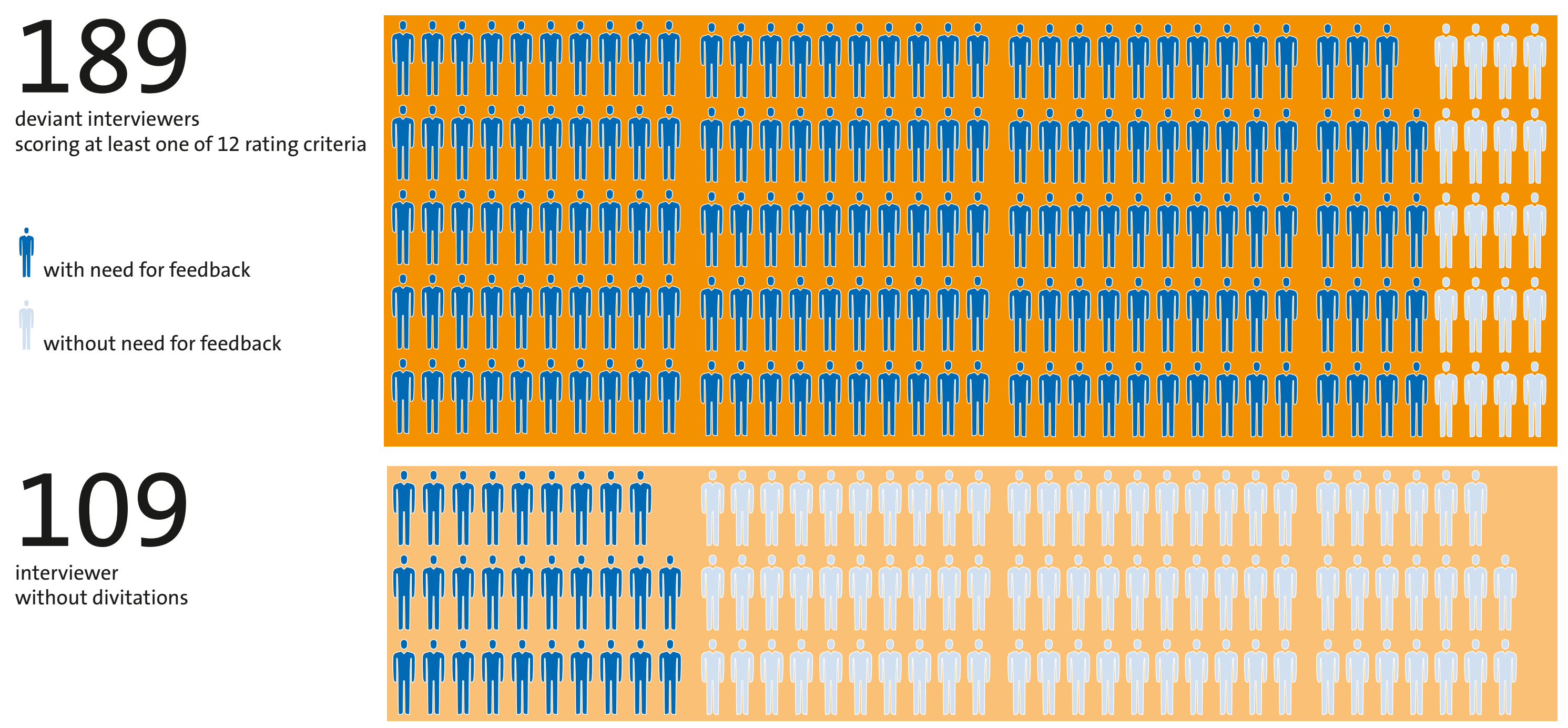
The 749 fully rated interviews refer to 298 interviewers. With the aim of monitoring interviewer behavior promptly at the beginning of the interviewer's work and providing quick feedback, the selection of rated interviews includes an average of 2.5 recordings per interviewer (maximum 8, minimum 1). An index based on the sum of criteria with deviations on the coding scheme was created for each interviewer. Every interviewer scored an average of one to two out of 12 rating criteria with deviations. 109 interviewers out of 298 fully rated interviews showed no deviations and 189 at least one deviation in any of the 12 rating criteria.

A total of 198 interviewers were identified as needing further training. Feedback and follow-up trainings took place by telephone and were conducted by the staff who had also edited the audio files and carried out the ratings. Further training should happen as soon as possible after the rating. Coding the audio recordings with the existing scheme was very time-consuming. Editing the audio recordings and documenting the deviations took an additional 15 to 30 minutes on top of the actual duration of the interview.
The feedback calls took additional 15 minutes. The whole effort for rated cases is on average about one hour (750 hours for 749 fully rated interviews, which means 2.5 hours per interviewer).

TABLE 1 Rating criteria and observed deviations (UIB) for fully rated interviews

Observed deviations (UIB) *Based on 749 audio recordings with full rating*		Observed per interview	Percentage of interviews n=749	Maximum observed per interview
Asking questions (standardized)	Not completely as presented	102	13.6	14
	Without all answer categories	73	9.7	5
	Without text accentuation	1	0.1	1
	Without necessary adaptions	2	0.3	1
Probing and clarifying	Incorrect reply to R's question	15	2.0	1
	With unpermitted explanation	65	8.7	6
	Missed necessary explanation	46	6.1	3
	Missed clarification	19	2.5	2
	Without active listening	18	2.4	2
Coding answers	Without correction of further answers	19	2.5	1
	Before clear matching of answer	130	17.4	19
	Being suggestive	228	30.4	38
Total		**370**	**49.4**	**39**

FIGURE 2 Statistical values observed UIB and need for feedback for fully rated interviewers



STATISTICAL PROCEDURES – ASSUMED UIB

Undesirable interviewer behavior can also be supervised by statistical methods. Deviant behavior in this sense assumes that the interviewer's behavior influences the respondent's answers and thus the data. The ICC measures the effect of interviewer behavior on distribution (moments of distribution, namely mean and variance). We performed an analysis in PASS based on the ICC after the end of field work in order to be able to make a comparison with the rating results.

Due to the behavior coding, two question modules could be identified in the household and in the personal interview, which were most frequently mentioned in the examples for the observed deviations. Overall, the calculation of the ICC for all items within the questionnaire modules indicates low interviewer effects for the individual items (below a level of 0.05 – see Table 2). A few indications of stronger effects can be observed, which then also exceed 0.10 for the coefficient. The module "Networks" is particularly prone to interviewer effects, especially the questions dealing with counting the number of persons with certain characteristics from the personal network. During PASS interviewer trainings these questions were also repeatedly reported as requiring explanation.

Conspicuous interviewers were identified and marked separately for each item. An index based on the sum of markings was created for each interviewer (referred to as "UIB assumed").

TABLE 2 Statistical values of assumed UIB for fully rated interviewers on questionnaire modules

Deprivation module 23 items
Deprivation module 4,954 households

Variable	ICC*	Item
HLS0300a	0.059	apartment with bathroom
HLS0800a	0.067	car
HLS0900a	0.065	television
HLS1000a	0.058	video recorder/DVD player
HLS1900a	0.118	going to the cinema/theatre/concert
HLS2200a	0.061	unexpected expenses with one's money
HLS2300a	0.069	medical treatment not fully covered
HLS2400a	0.058	rent payment for apartment on time

HLS0100a, HLS0200a, HLS0400a, HLS0600a, HLS0700a, HLS1100a, HLS1200a, HLS1400a, HLS1500a, HLS1600a, HLS1700a, HLS1800a, HLS2000a, HLS2100a with an ICC between 0.014 and 0.049

Networks module 21 items
Social network module 8,074 persons

Variable	ICC*	Item
PSK0280b	0.068	somebody who tells about vacant job
PSK0280e	0.064	somebody who helps with job application
PSK0280f	0.063	somebody who recommends you to an employer
PSK290a	0.090	number of close friends with high school degree
PSK290b	0.155	number of close friends without education degree
PSK290c	0.149	number of close friends unemployed
PSK290d	0.131	number of close friends with 'Minijob'
PSK290e	0.118	number of close friends self-employed
PSK0300	0.055	misunderstandings in household
PSK0500	0.080	time per week for voluntary activities
PSK0600a	0.055	going out with friends
PSK0600f	0.053	going on trips with friends

PSK0100, PSK0200, PSK0280c, PSK290f, PSK290g, PSK0600b, PSK0600c, PSK0600d, PSK0600e with an ICC between 0.009 and 0.049

**ICC significant over all items on level 0.05*

UIB: OBERSERVED AGAINST ASSUMED

Comparing the results from the statistical calculation (assumed UIB) with the rating results (observed UIB) is possible on the interviewer level by using the two indices. The statistical calculations revealed no abnormalities for 140 interviewers in the deprivation module and 89 in the network module. Overall, it can be concluded that the results from behavior coding and multivariate analyses complement each other so that interviewers with poor overall rating also show systematic effects on response destributions. This results in an overall impression that behavior coding identifies other interviewers or other aspects of behavior and is less recognizable by statistical calculations. In contrast, the statistical method emphasizes more systematic deviations more clearly.

TABLE 3 UIB for fully rated interviewers: statistical measurement (assumed UIB) compared to behavior coding (observed UIB)

Interviewers Assumed UIB *Interviewers with full rating only.* *Suspicious interviewers' ICC analysis see table 2* *Deviation on 12 rating criteria see table 3*	**Without observed UIB**		**With observed UIB**		**Total**	
	Obs	*Percent*	*Obs*	*Percent*	*Obs*	*Percent*
Deprivation module						
0 items	46	42.2	94	49.7	**140**	**47.0**
1 to 8 items	43	39.5	74	39.2	**117**	**39.3**
9 and more items	20	18.4	21	11.1	**41**	**13.8**
Networks module						
0 items	31	28.4	58	30.7	**89**	**29.9**
1 to 11 items	65	59.6	114	60.3	**179**	**60.1**
12 and more items	13	11.9	17	9.0	**30**	**10.1**
Total per module	**109**	**100.0**	**189**	**100.0**	**298**	**100.0**

RESULT

The behavior coding clearly showed that interviewers deviated more frequently while dealing with the respondents' answers than while reading out the question and answer categories. The concrete examples from the ratings also hint at questions in the questionnaire, in which interviewers often act undesirably. This advantage is offset by the costs of the process. Monitoring in the CAPI field through audio recordings could potentially be even more effective with a shorter rating scheme. In addition, reducing costs would result from written feedback. Follow-up trainings could then be targeted specifically to special individual cases. infas is already preparing suggestions on further ideas for optimizing this procedure, which are to be tested in the context of the upcoming PASS Wave 13. Comparing both methods – behavior coding and statistical measurement – has shown that the results are complementary. Interviewers with deviant behavior are well identified with both methods. It certainly plays a role that the behavior coding takes into account a small part of interviewer behavior only. In addition, the calculation may include all interviewers deployed and not just those for whom audio recordings are available. However, statistical methods can only fall back on a substantiated database if a sufficiently large number of cases is available. Unfortunately, this is always the case at a later stage in the field. However, the effort with regard to time and costs is much less than for behavior coding because it does neither depend on case numbers nor study design. In addition to behavior coding, initial statistical calculations could actually be carried out earlier in the field. Indications of possible interviewer effects could be the trigger for the targeted rating of audio recordings of specific interviewers and thus support behavior coding.

For effective monitoring in the CAPI field, we advocate a combination of both procedures, which tooks place during the whole field period and could reduce one source of survey errors.

FIGURE 3 Survey life cycle for CAPI-interviewer with statistical procedure only

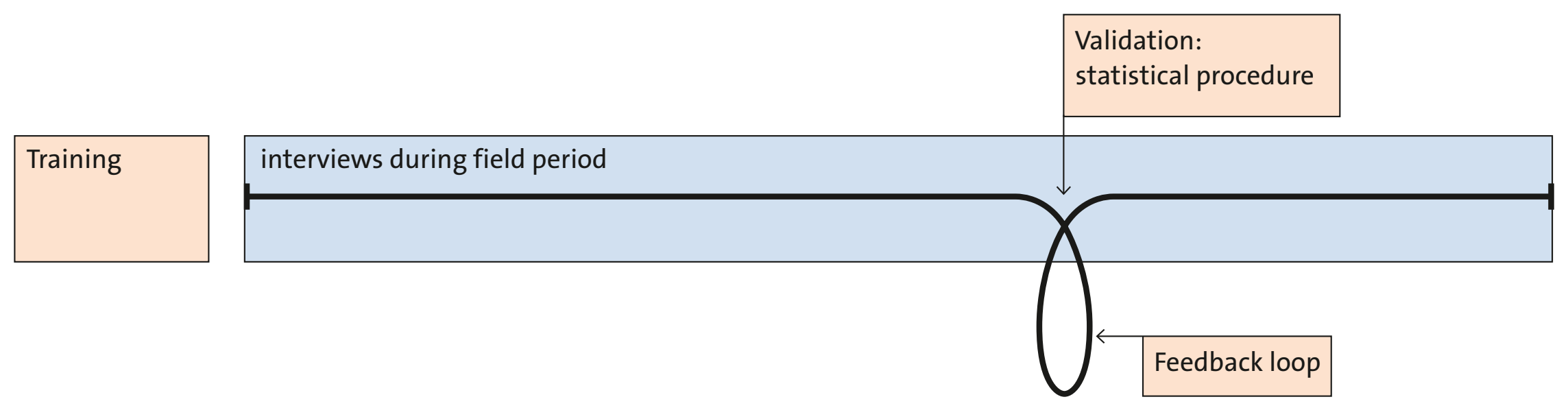
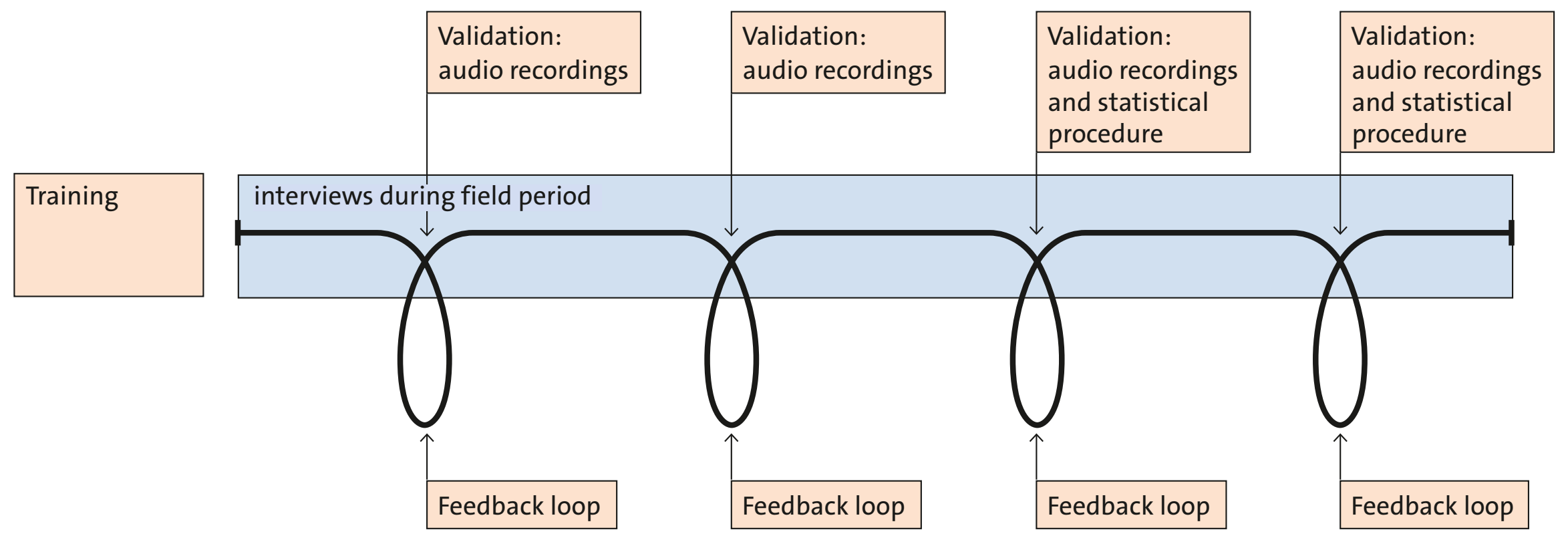


FIGURE 4 Survey life cycle for CAPI-interviewer with combination of audio recordings and statistical procedure



International Field Directors & Technologies Conference, Interviewers and Their Effects from Total Survey Error Perspective Workshop, 2019, Poster Session "Case Assignment and Monitoring"