MASARYK UNIVERSITY
FAKULTY OF SOCIAL STUDIES

Quantitative Methods on the Internet: Experiences from Online Surveys

prof. David Šmahel, Ph.D.
Hana Machackova, Ph.D.

Methods: offline and online

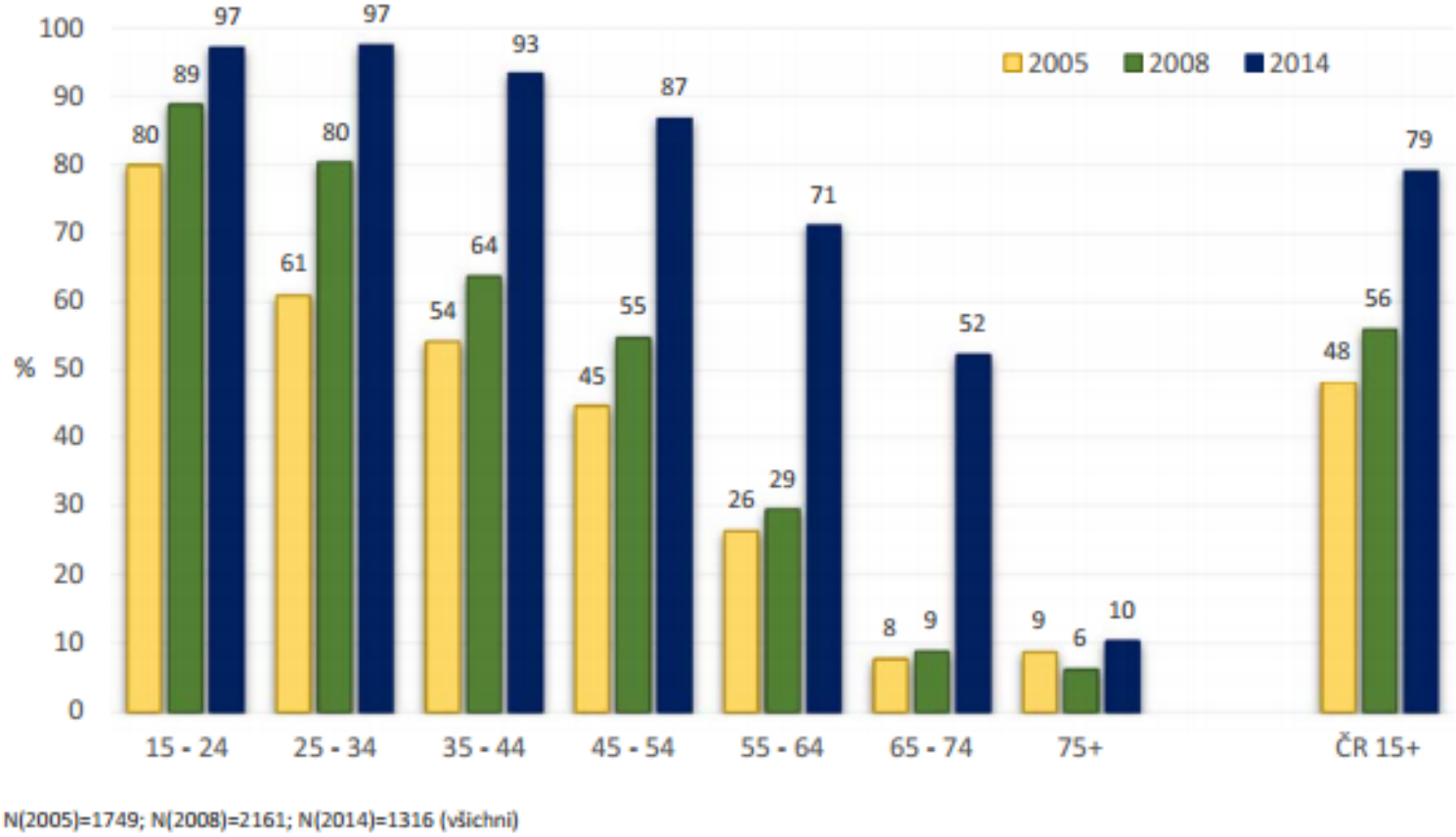
- Qualitative
- Quantitative (online)
- Content analyses
- Observation / (n)etnography
- Metaanalyses
- „Technical “ methods
- Mixed methods

Table 10.1 CASIC modes according to interviewer involvement

CASIC mode	*Interviewer involvement*	*Brief description*
CATI – *Computer-assisted telephone interviewing*	Remotely present	The first CASIC mode. An interviewer calls respondents by phone and enters answers into the computerized questionnaire.
CAPI – *Computer-assisted personal interviewing*	Physically present	The mode enabled by introduction of portable computers. An interviewer brings a portable computer with the questionnaire to respondents and enters answers into it.
CASI – *Computer-assisted self-interviewing*, Audio-CASI, Video-CASI	Physically present	Similar to CAPI but respondents answer the questionnaire on an interviewer's computer by themselves. Variations are audio-CASI and video-CASI, where questions are presented using audio or video clips.
CAVI – *Computer-assisted video interviewing*	Remotely present	Similar to CATI but the communication between an interviewer and respondents is established using video calls or similar technology.
Disk-by-mail	Not present (CSAQ)	Respondents answer – using their own computer – the questionnaire on a floppy disk sent by the researcher.
TDE – *Touch-tone data entry*	Not present (CSAQ)	Respondents input their answers by pressing appropriate numeric keys on a telephone handset.
IVR – *Interactive voice response*	Not present (CSAQ)	A wide range of approaches for voice communication with a computer system using the telephone. Modern IVR systems, supported by speech-recognition technologies, already enable respondents to provide complex answers through the telephone that are automatically recorded as text.
Internet surveys	Not present (CSAQ)	A variety of survey modes in which questionnaires are delivered and answered using Internet technology (e.g. e-mail or web). The most widely used are web surveys and less used e-mail surveys.
Virtual interviewer surveys	Not present (CSAQ)	Questions are presented to respondents using some kind of virtual interviewer, usually through the Internet. Future technological development will enable increased virtualization of the surveying process, where interviewers will probably become completely computerized virtual characters.

(Vehovar & Manfreda, 2008)

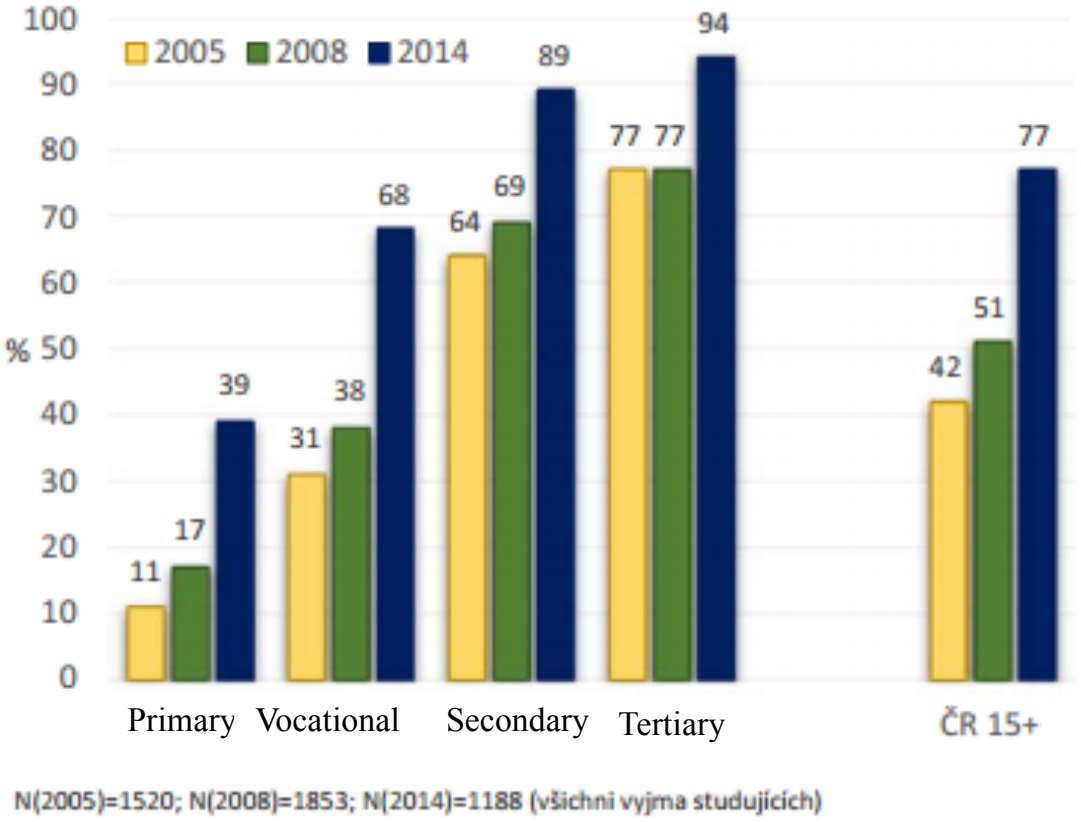
Internet users: the Czech Republic



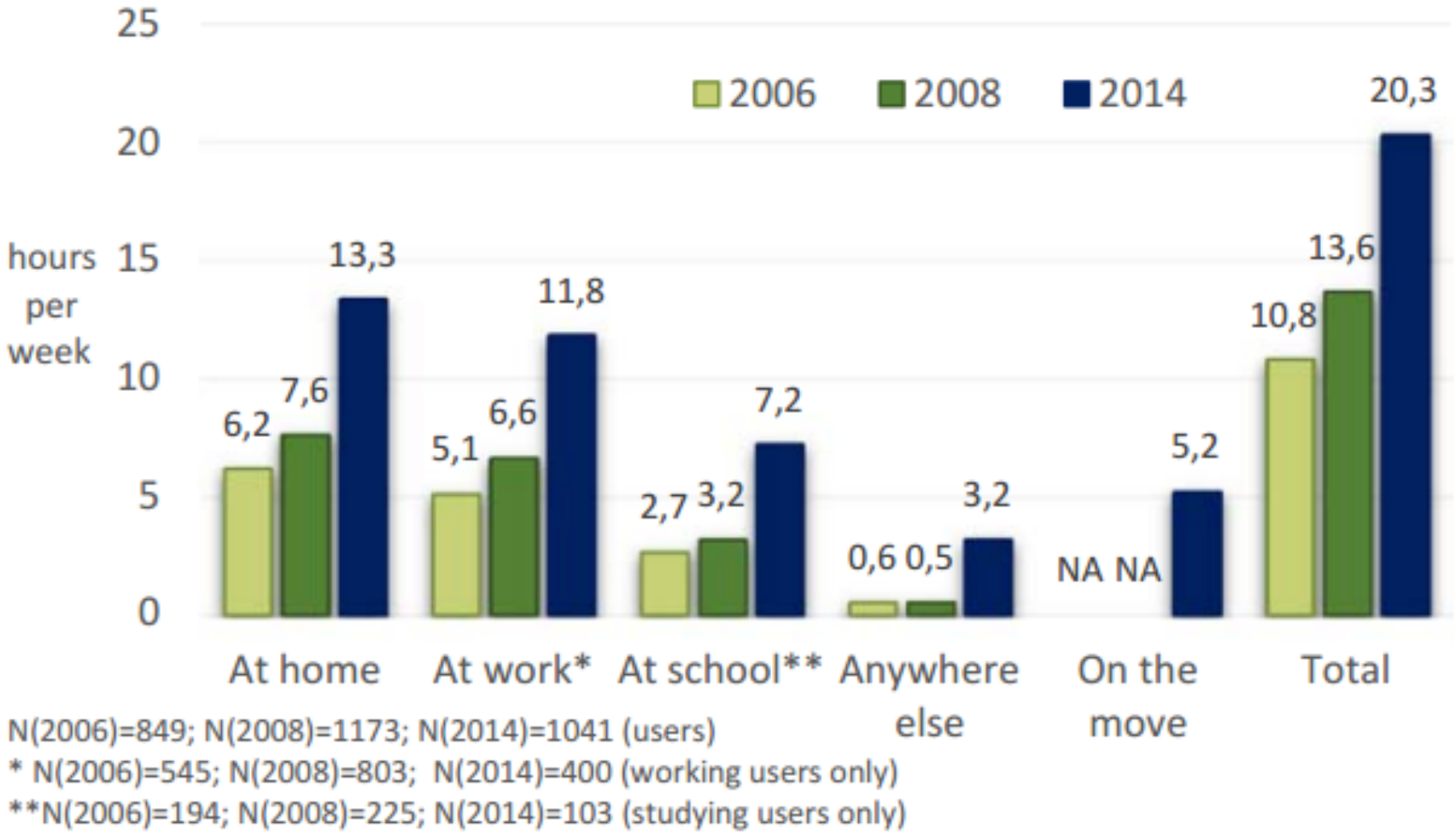
N(2005)=1749; N(2008)=2161; N(2014)=1316 (všichni)

(World Internet Project – Lupač, Chrobáková, Sládek, 2014)

Internet users: Education



N(2005)=1520; N(2008)=1853; N(2014)=1188 (všichni vyjma studujících)



(World Internet Project – Lupač, Chrobáková, Sládek, 2015)

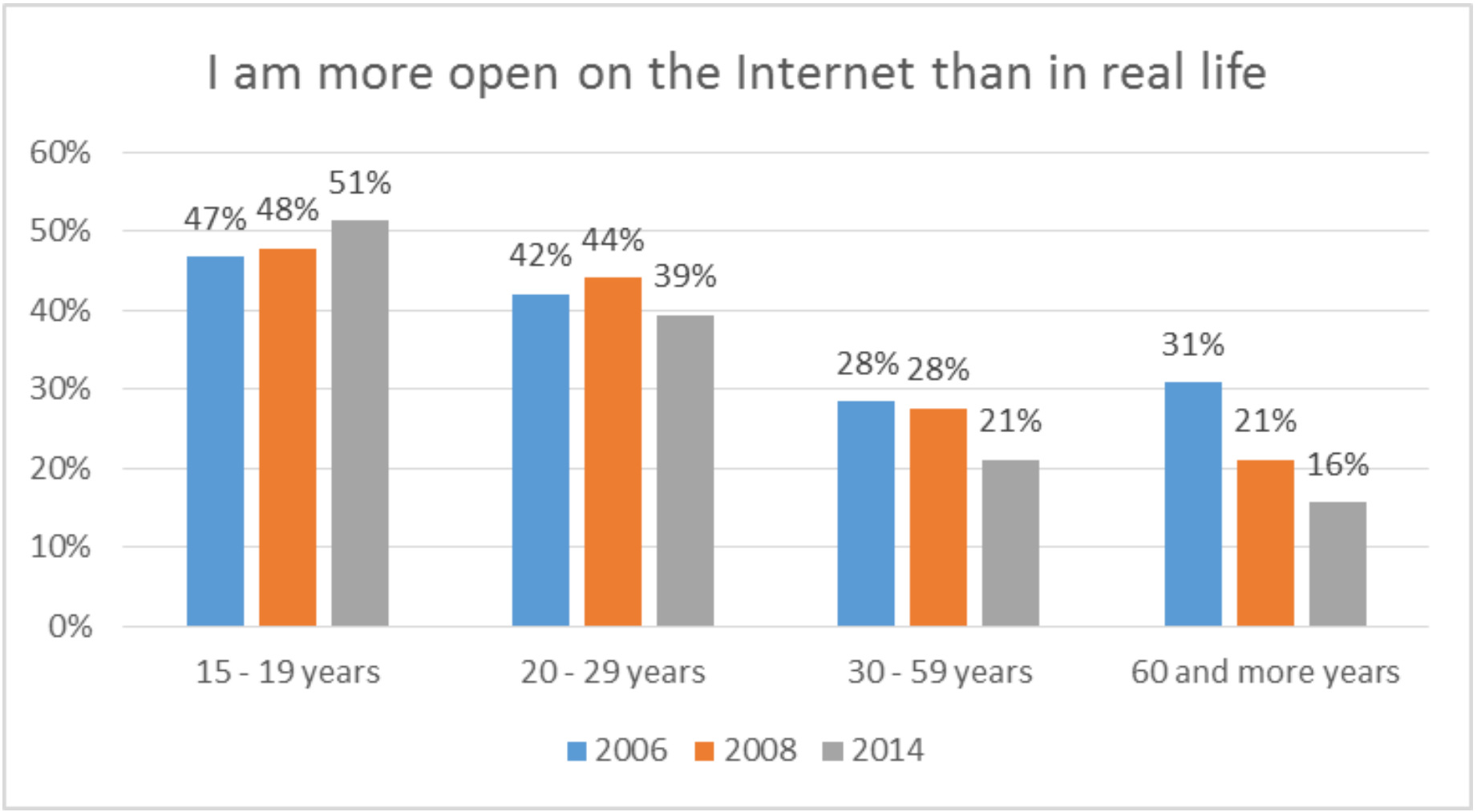
Internet population

- Internet population is not a representative sample in any country
- It has specific characteristics
- We do not know the population differences from the psychological point of view
- Who is in our sample?
  - We cannot check even gender and age
- Parallel with the critic of the research carried on university students

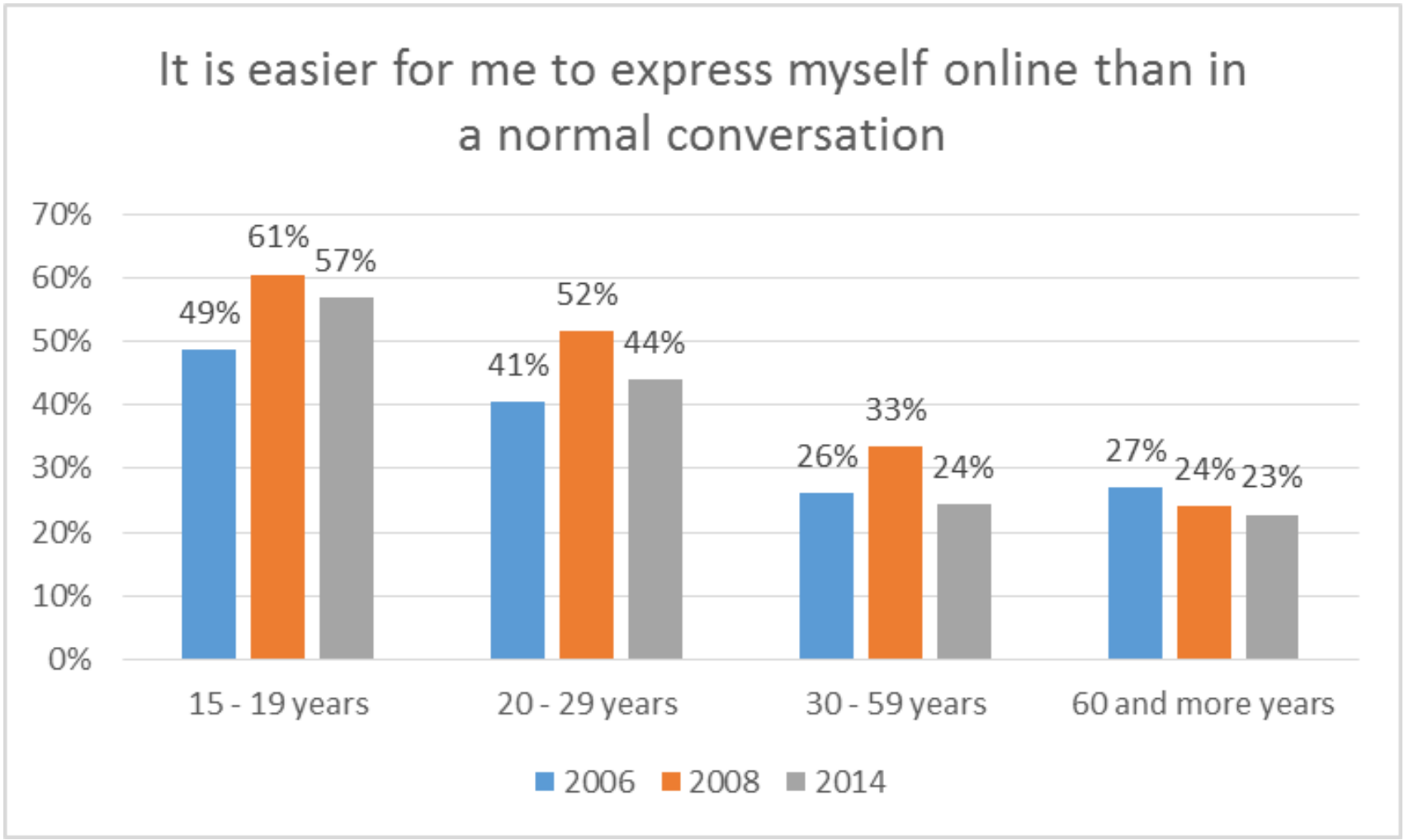
Internet as a medium

Suitable for research into:

- specific groups (students, organisations, IT professionals, scientists, etc.)
- subcultures otherwise difficult to reach (drug community, hackers, sexual deviations, etc.)
- communities appearing on the Internet (chat rooms – e.g. homosexual dating services, blogs, interest groups, etc.)
- ‘sensitive issues’ – higher openness due to anonymity



(World Internet Project; Šmahel, 2015 conference Virtuálna generácia, 2015, Bratislava)



(World Internet Project; Šmahel, 2015 conference Virtuálna generácia, 2015, Bratislava)

Where we can collect data online?

The choice is based on the selection of the given population:

- Web pages – various groups
- Social networks
- Blogs
- Chat rooms
- Discussion ‘boards’
- Online games
- Accessible email databases (schools, or organisations)
  - careful when consent acquirement needs to be addressed!
- Instant messengers (e.g. Skype, Facebook…)

Ethical problems

- Various approaches of ethical boards to online research
- How to research children when we need consent of parents?
- In some countries, the ethical approval is not needed from law perspective in strictly anonymous research (when “personal data” are not collected)
- How to get consent from respondents?
- How to get approval from parents?
- Problems of data safety
- Privacy in online surveys

How to address respondents for the survey

E-mail:

- the most invasive method but most powerful
- recommended to address with personification (if we have data for that)
- recommended to keep recipients' attention
- there are the same rules as in real questionnaires
- rate of the questionnaire return is 2-50% (!)
- be careful of SPAMing

How to address respondents for the survey

Searching respondents in chat rooms and/or web discussion boards

- has more characteristics of an opinion poll
- it depends on the environment and research type
- the rate of the questionnaire return is debatable

Personal addressing through messengers

- pretty effective but laborious (addressing always only 1 respondent at a time)

How to address respondents for the survey

Social networks

- very popular and easy
- sampling isues – addressing mostly specific subpopulations
- depends on the particular SNS and where we start („networks“)

Advertisement on www (like pop-up window, banner)

- the poorest method
- mostly the character of an opinion poll,
- click rate of banners is 0.1 - 2%

How to collect data

Best web questionnaire

- professional x free services (i.e. Limesurvey), broad possibilities, adaptive testing, collecting a lot of metadata (IP addresses, measuring time in the questionnaire), export in databases

Sending questionnaire in an Excel file or Word

- less suitable, part of respondents will not download it, will not open it etc.

Text e-mail

- not providing an easy survey, very hard data processing

Administrated survey (e.g., via skype)

- some benefits of administrated data collection (more control, providing explanations , etc.)
- less anonymity, much more demanding

Non-response problem & Motivating respondents

- Will you allow non-response of items in the survey?
  - Yes -> missing data
  - No -> low response rate and more bias
- Which items will you select as “must-fill”?
- Response rate in online surveys: 1 – 100%
- Motivation of respondents is crucial
- What is the right motivation? For your particular targeted population?
- How to provide rewards and assure anonymity?

Factors influencing answers of your respondents

- Motivation (!!)
- Digital literacy
- Attitudes of the user to the digital technology
- Privacy issues
- Design of the online survey
- PLUS what we know from offline surveys: i.e. self-presentation, problems of answering sensitive questions etc.

(See also: Vehovar & Manfreda, 2008)

The data processing

BEFORE - checking the functionality

- design, saving of the correct data, filtering

AFTER

- Cleaning data
  - using also metadata
  - anonymization
- Checking the all attributes
  - distribution, reliability analyses, factor structure, etc.
- Checking the sample
  - and possibilities of generalization

Conclusion: Positives and Negatives of Online Surveys

+ Sample extent, global reach, cost reduction, time saving, preserving anonymity, respondents' greater openness, access to specific populations, survey participants' comfort, minimising interviewer-related bias, and research methods flexibility, longitudinal research might be cheap.

- Non-representative sample, possibly distorted replies (higher possibility of lies and hypocrisy), loss of information about the research process context, limiting non-verbal elements of communication, technical possibilities of the researcher and respondent, and absence of direct contact.

Our experiences: research on the users of websites focused on nutrition, diet, or fitness

Sampling and data processing

…and general preparation of the online survey

Sampling process

A very different platforms….

4 researchers – creating **shared database** of Czech online websites devoted to the topic

Need to have all **important information about site:**

- URL
- Name of the site
- Type of web
- Where to announce (are there discussions? Advertisement?)
- Contact (whom? In what form?)
- How visited the site is (not always reliable)
- Connected to FB or something else?

Sampling process

A very different platforms….

4 researchers – creating **shared database** of Czech online websites devoted to the topic

Need to have all **important information for progress and cooperation:**

- Who found contact? Who innitiated contact?
- When was the site contacted?
- Was there reaction?
- If positive, was the invitation really published?

Sampling process

The contact:

Need for repeated contact and checks

Contact via different channels (email, FB, phone)

Type of contact

- Mostly via email – not really effective
- Via Facebook – similar
- More direct – telephone
  - demanding – contacted only big sites
  - very good results (could depend on the type of research)

- In total, 307 sites were reached, most of them did not write back
- Only 49 agreed and published the invitation

Sampling process

Clustered sampling – need to address both „clusters“ (site owners) and respondents

Several types of contact forms and invitations, depending on…

- the type of invitation – short article, banner, facebook text…
- online platform – email vs. facebook

Similarities:

- Research institution – official
- Importance of research and its implications
- Motivation
- Population (web visitors, age range)

Sampling process

Clustered sampling – need to address both „clusters“ (site owners) and respondents

Several types of contact forms and invitations, depending on…

- the type of invitation – short article, banner, facebook text…
- online platform – email vs. facebook

Changes:

- Length
- Formal/informal language

Sampling process

Používáte internet nebo telefon k hledání informací či názorů a zkušeností ostatních ohledně stravování, cvičení či sportování? Například o zdravé stravě, vhodné dietě, nebo informace ohledně udržení či vybudování kondice?

Pokud ano, chtěli bychom Vás požádat o pomoc s probíhajícím výzkumem Masarykovy univerzity podpořeným Grantovou agenturou České republiky zaměřeným na mladé lidi (ve věku 13-28 let). Pokud nám vyplníte online dotazník, pomůžete nám pochopit, jak lze používat technologie v kontextu stravování a cvičení.

Dotazník je anonymní a zabere asi 20 minut. Jako poděkování za Váš čas a ochotu se můžete zúčastnit slosování o 5 poukázek v hodnotě 1000 Kč na nákup na Mall.cz. Po skončení výzkumu se také můžete podívat na základní výsledky na stránce http://thinline.fss.muni.cz/o-projektu/kvantitativni-cast.

Odkaz na dotazník:
http://www.urbandevelopment.cz/research/index.php/926229/lang-cs?926229X171X5051=2984

S velkým díky za případnou pomoc
Prof. PhDr. David Šmahel, Ph.D. (řešitel projektu)

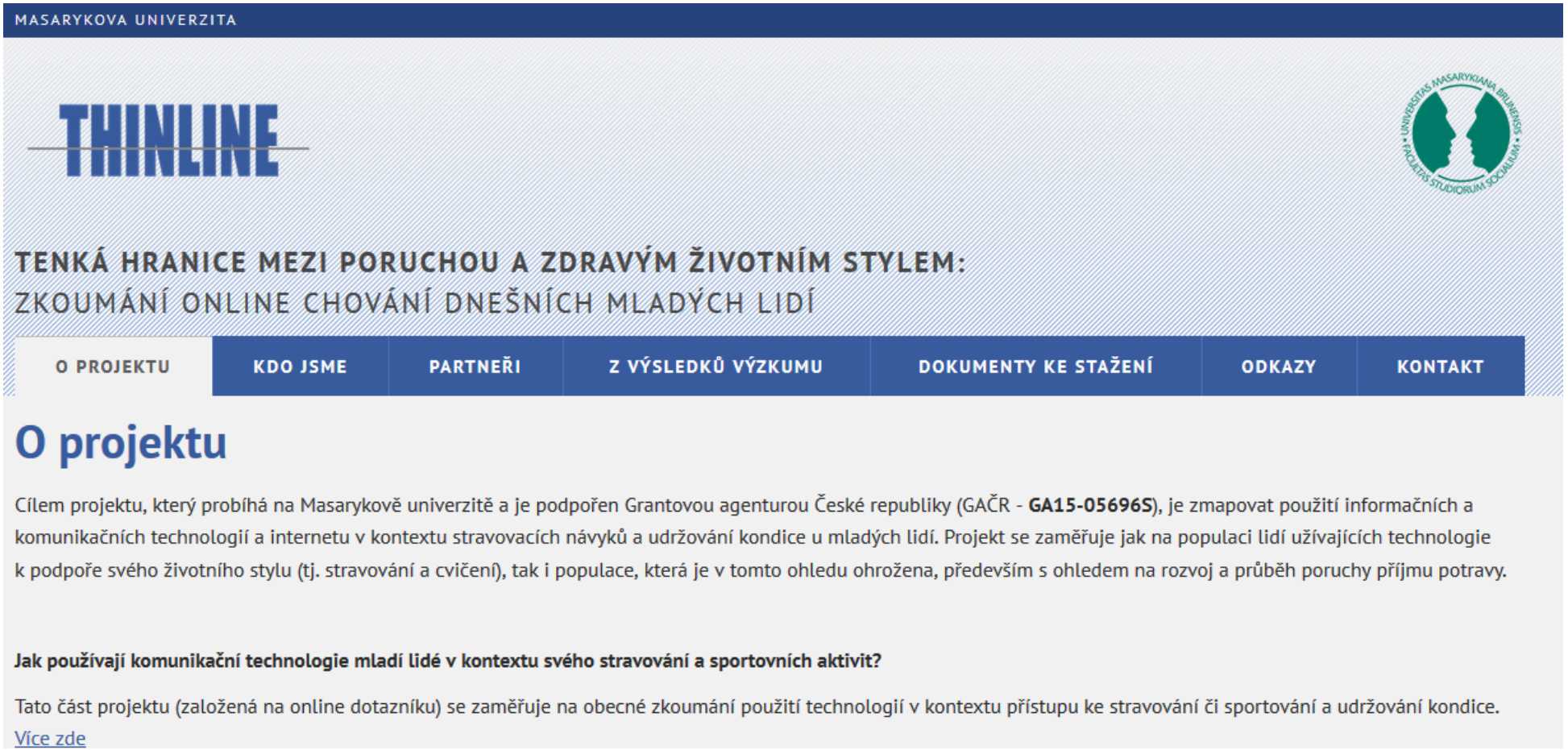
Institut výzkumu dětí, mládeže a rodiny
Fakulta sociálních studií
Masarykova Univerzita Brno

Sampling process

Sampling process

Our own website and FB page

- More details
- Connection to University (official address)



Sampling process

Motivation. Clustered sampling!

- Most people do not like questionnaires
- Most web owners do not want to bother their visitors with questionnaires

Two types for motivation

„Implicit“

- Emphasize the contribution to knowledge (but…)
- Sharing results (report) – adminstrators often liked it

„Explicit“

- Publishing cooperation on the website
- Chance to win a voucher

Sampling process

We planned the survey from May till end of June

- Aim 1,000 respondents

We collected data till the end of October

Very slow proces

- Delays in communication with the site owners

Peaks in data – annoucements (first page, top)

Quick decline - „old news“

Re-negotiations about making it „fresh“

Depending on the type of site and used invitation

We had data in the end of October…

A little step back…

- What we need? What we need to check?
  - To have reliable and valid data
- Several issues which need to be considered

- Upon these should be selected most suitable survey program
- we used Limesurvey

A little step back…

- **Mandatory questions**
  - Pros and cons (Davids‘ talk)
  - We tried to minimize them
- What we <u>really</u> needed to know?
  - Gender and age
  - Data about visits of websites
- Data needed for filtering
  - E.g., use of smartphone – use of smartphone apps

A little step back…

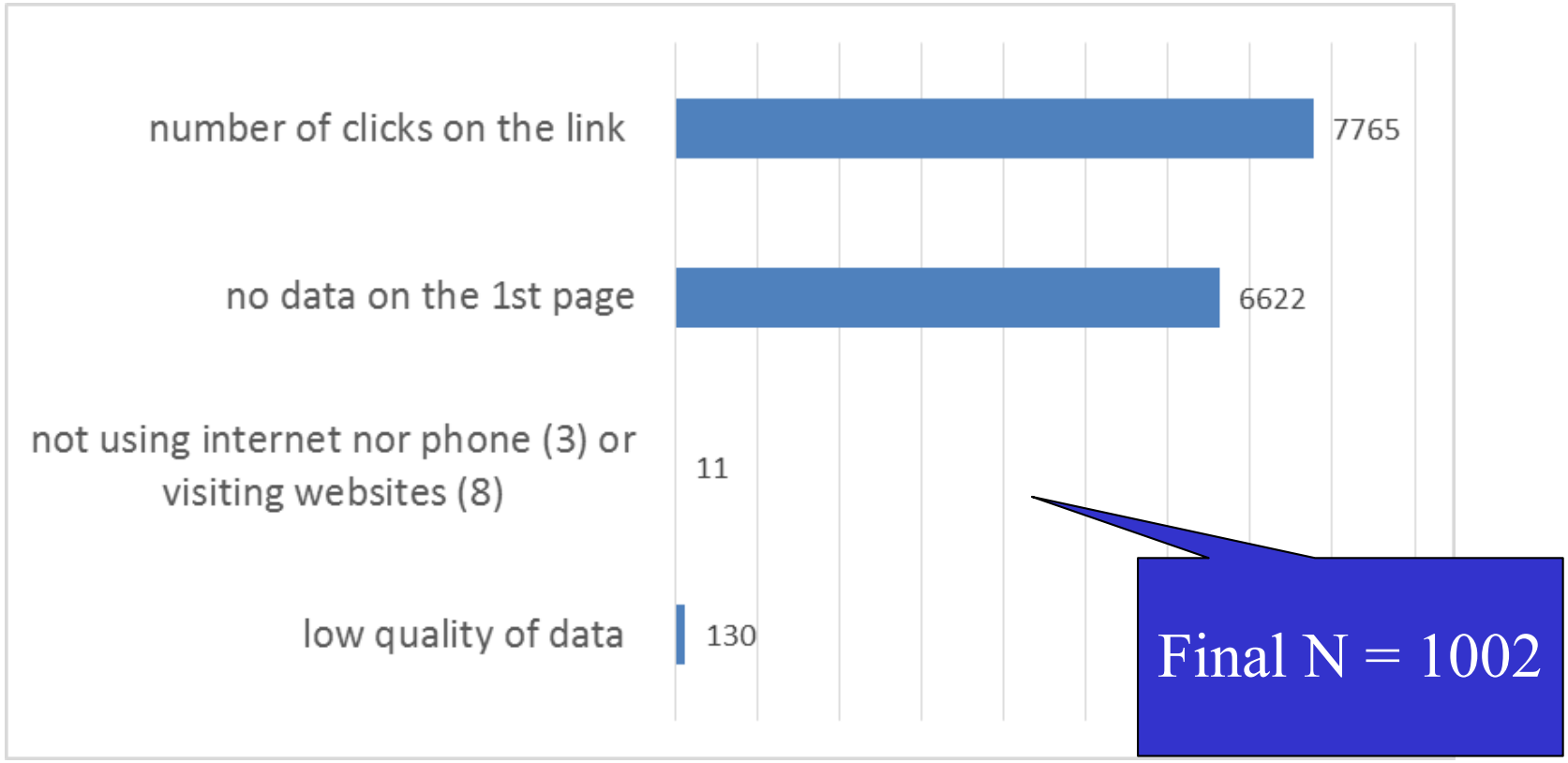
- **Filtering questions**
  - Several checkups how they work
  - Usually there is an error☺
- If you have an IT guy for survey adminitration
  - be ridiculously specific
  - they do not read the items, only instructions how to implemen them
  - do not assume „common sense“ (for social science research)
- Regular checks on the data and a lot of backups

A little step back…

- **Metadata** – what we need for data cleaning?
  - Timestamp (start, end)
    - For us not so important, but could be
    - E.g., research on politics and elections
  - time spent on whole Q and pages
  - IP address
  - URL (redirected)
    - Where did they come from?

Data processing

- The data always need to be cleaned



Data processing

The basics are the same as in offline survey

Check the sample (first)

Check the measures and raw data

Consider lying, boredom, „fun“

Many versions of the dataset (often need to go back)

Data processing

1) We got rid of „nodata“

Mix of genuine looks and bots

- Clicks from website owners
- Clicks of those interested – but in the end not interested

Set a line – for us, no click on the items (blank Q)

This reduced N about 7 times

Data processing

2) Checking coding, values, numbers

Sometimes problems with…

- Transfering (numbers, text, symbols, parameters in SPSS…)
- Mistakes due to changes in Q
  - E.g., scale 1, 2, 4, 5, 6
- And some other „mysteries“

Data processing

3) Check the sample

Outliers – affect most attributes

- e.g. children

The invitation was specifically for 13-28

- About one fifth of the sample was out of this range (mostly older)
- What to do with the data?

Data processing

4) Checking errors and bias…

Response sets

We do not have the visual form (snakes in coding)

Basic checking on the propensity to give the same answer

Preparation:

Rotated or non-sensical items

- but even non-sense may make sense sometimes

Cross-checking with the same information elsewhere

Data processing

4) Checking errors and bias…

Response sets

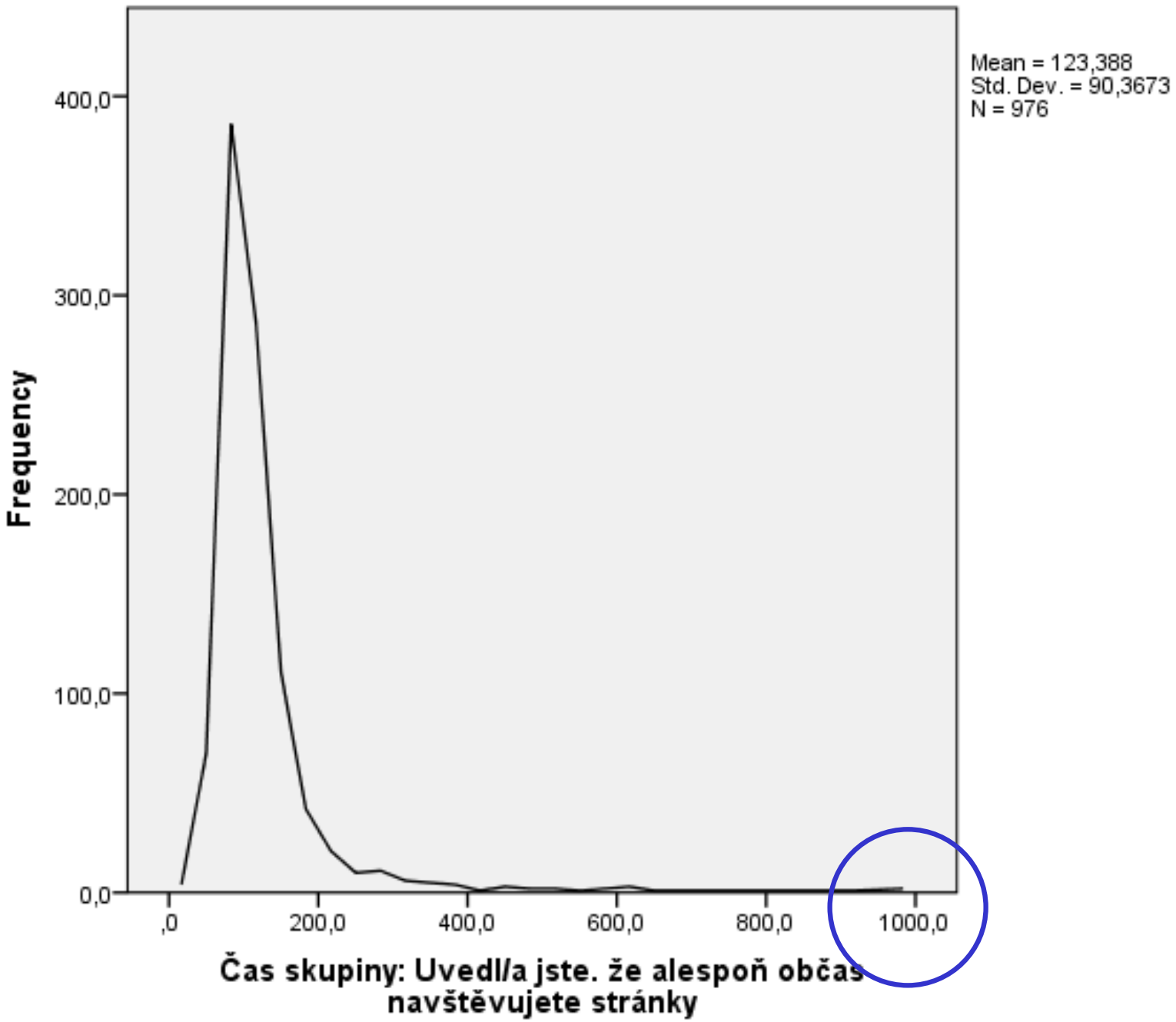
Also related to time spent on the items, scales, and pages

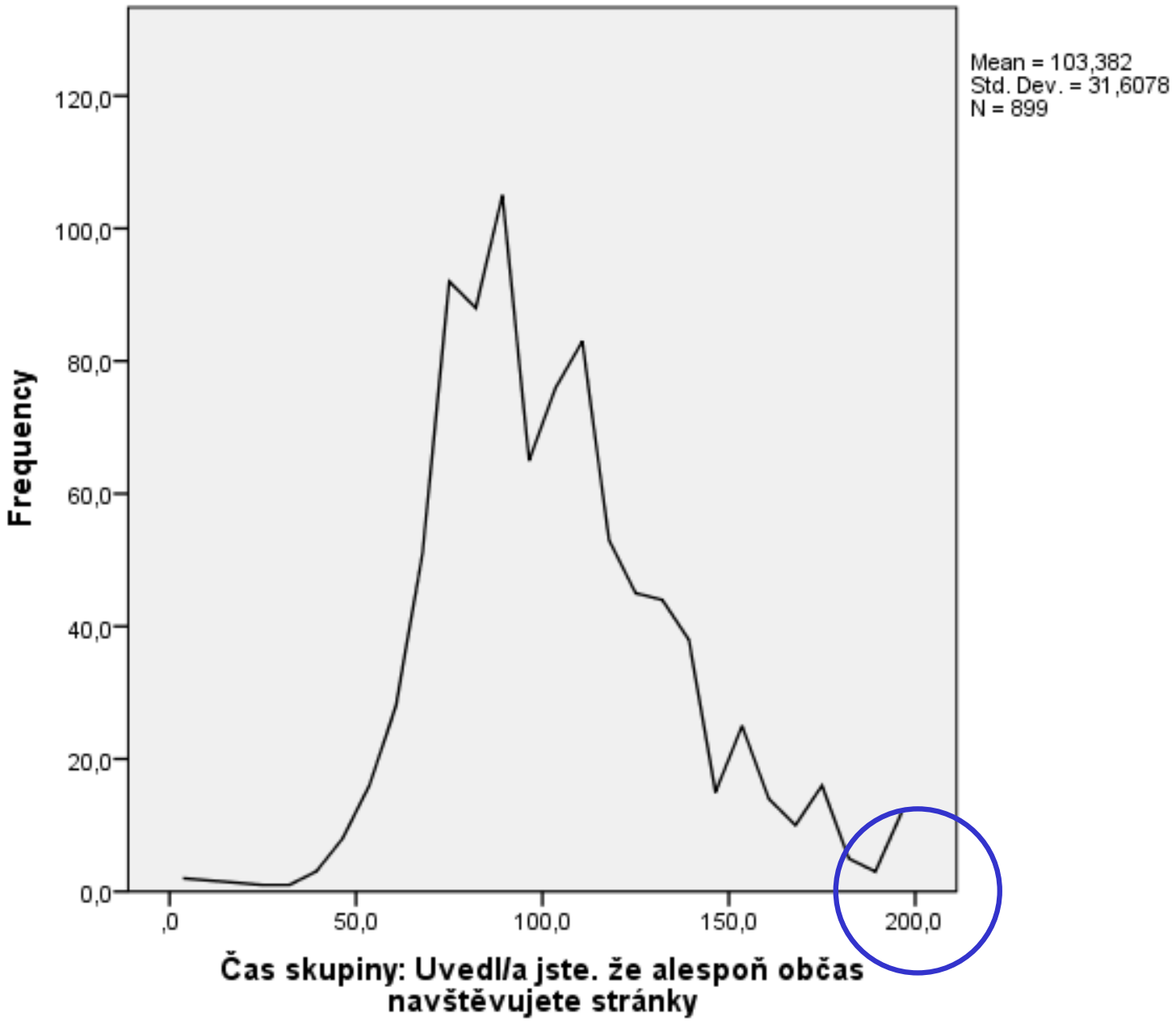
- good to structure them according to this expectation
- however, there is often no clear cut-off point
- no upper limit – open website even for hours

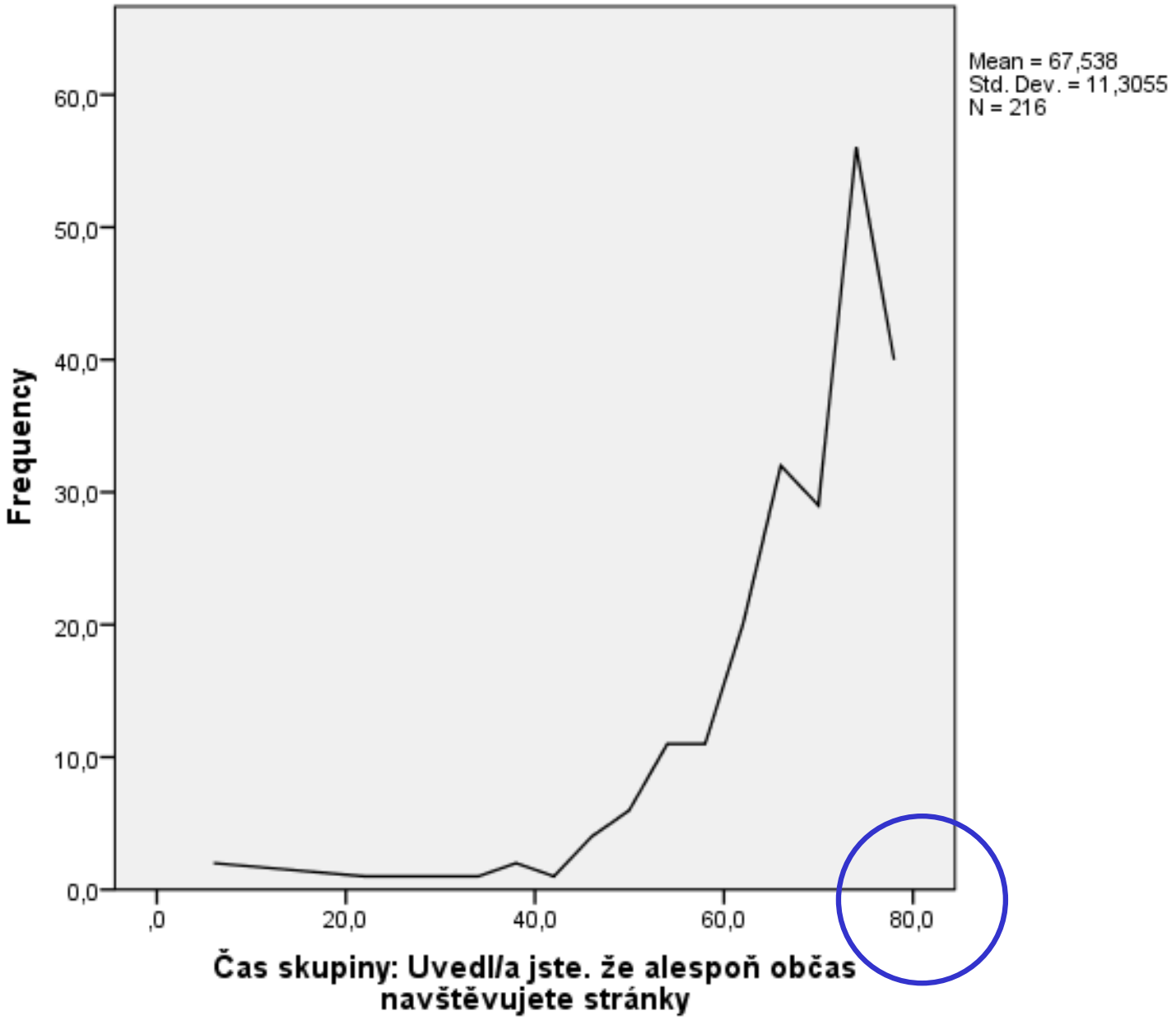
Time on p.2 (in seconds)

Čas skupiny: Stránky 2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	6,48	1	,1	,1	,1
	6,78	1	,1	,1	,2
	8,46	1	,1	,1	,3
	10,06	1	,1	,1	,4
	11,29	1	,1	,1	,6
	15,66	1	,1	,1	,7
	15,89	1	,1	,1	,8
	18,98	1	,1	,1	,9
	26,56	1	,1	,1	1,0
	28,57	1	,1	,1	1,1
	43,40	1	,1	,1	1,2
	46,46	1	,1	,1	1,3
	48,41	1	,1	,1	1,5
	50,38	1	,1	,1	1,6
	50,48	1	,1	,1	1,7
	52,16	1	,1	,1	1,8
	52,55	1	,1	,1	1,9
	52,57	1	,1	,1	2,0
	53,19	1	,1	,1	2,1
	53,29	1	,1	,1	2,2
	53,70	1	,1	,1	2,4
	53,88	1	,1	,1	2,5
	54,08	1	,1	,1	2,6
	55,06	1	,1	,1	2,7
	55,11	1	,1	,1	2,8
	376,78	1	,1	,1	97,3
	378,90	1	,1	,1	97,4
	386,97	1	,1	,1	97,5
	446,69	1	,1	,1	97,6
	446,87	1	,1	,1	97,8
	469,63	1	,1	,1	97,9
	476,92	1	,1	,1	98,0
	497,18	1	,1	,1	98,1
	525,54	1	,1	,1	98,2
	547,07	1	,1	,1	98,3
	581,08	1	,1	,1	98,4
	590,79	1	,1	,1	98,5
	633,07	1	,1	,1	98,7
	667,14	1	,1	,1	98,8
	732,46	1	,1	,1	98,9
	757,25	1	,1	,1	99,0
	821,14	1	,1	,1	99,1
	1079,21	1	,1	,1	99,2
	1318,01	1	,1	,1	99,3
	1390,76	1	,1	,1	99,4
	1626,78	1	,1	,1	99,6
	1677,43	1	,1	,1	99,7
	2020,43	1	,1	,1	99,8
	2542,50	1	,1	,1	99,9
	3336,63	1	,1	,1	100,0
	Total	891	82,1	100,0	
Missing	System	194	17,9		
Total		1085	100,0		


Mean = 123,388
Std. Dev. = 90,3673
N = 976
Frequency
400,0
300,0
200,0
100,0
0,0
,0
200,0
400,0
600,0
800,0
1000,0
Čas skupiny: Uvedl/a jste. že alespoň občas navštěvujete stránky


Mean = 103,382
Std. Dev. = 31,6078
N = 899
Frequency
120,0
100,0
80,0
60,0
40,0
20,0
0,0
,0
50,0
100,0
150,0
200,0
Čas skupiny: Uvedl/a jste. že alespoň občas navštěvujete stránky


Mean = 67,538
Std. Dev. = 11,3055
N = 216
Frequency
60,0
50,0
40,0
30,0
20,0
10,0
0,0
,0
20,0
40,0
60,0
80,0
Čas skupiny: Uvedl/a jste. že alespoň občas navštěvujete stránky

Data processing

5) Checking the sample again and the measures in our sample

6) And any other additional info needed

Clustered sample – from which site they came?

3 types of info – written text (by respondents), re-directed URL, and links with specific code for every website