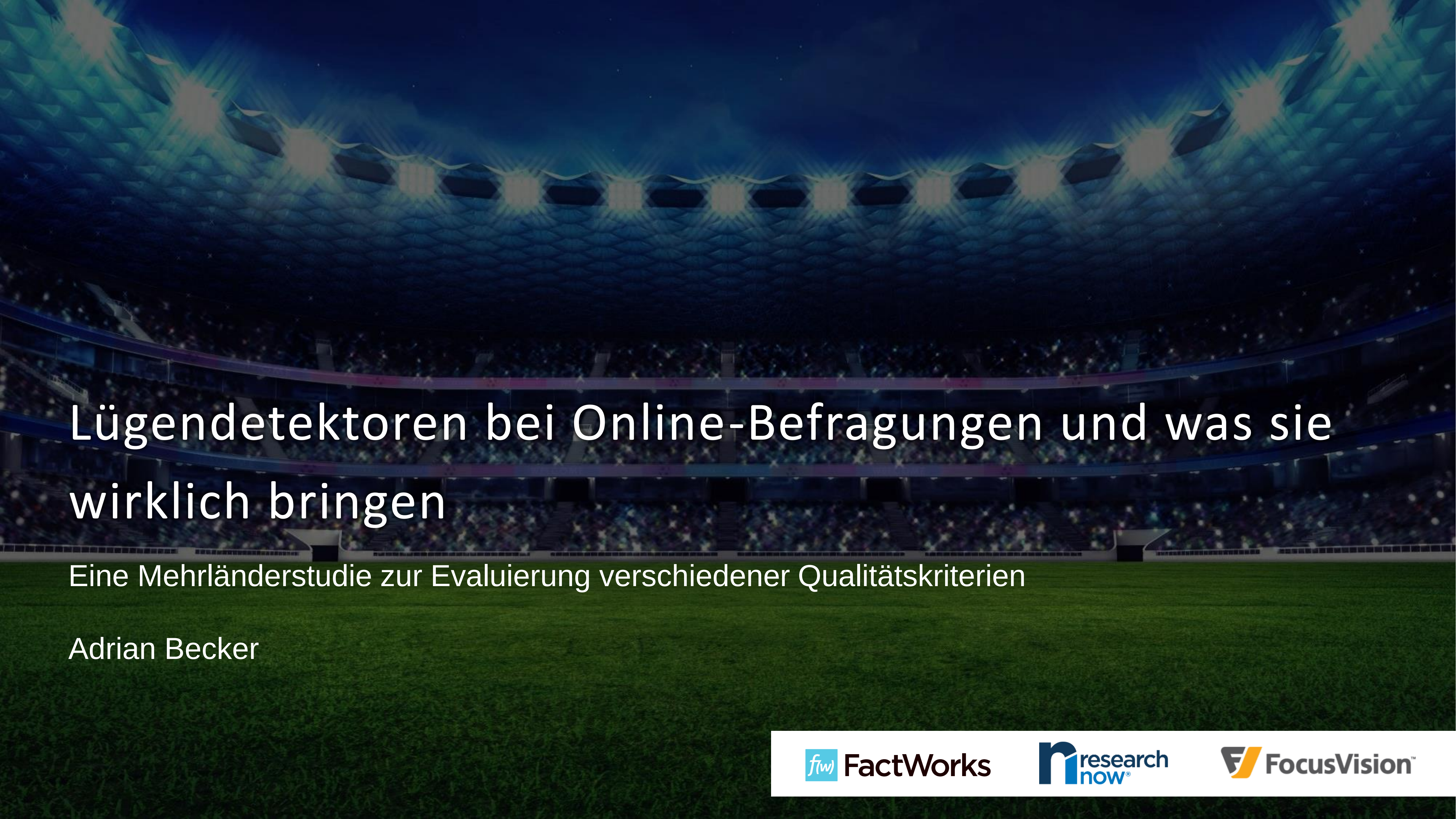


Lügendetektoren bei Online-Befragungen und was sie wirklich bringen

Eine Mehrländerstudie zur Evaluierung verschiedener Qualitätskriterien

Adrian Becker

FactWorks research now FocusVision

High data quality is a hygiene factor (must-have) in our business and is directly linked to delivering reliable, actionable information.

This study provides guidance on how to improve data quality

DRIVERS OF DATA QUALITY

Survey Design

Respondents behavior

Variety of methods to detect “bad”
(i.e., inattentive or fraudulent) respondents

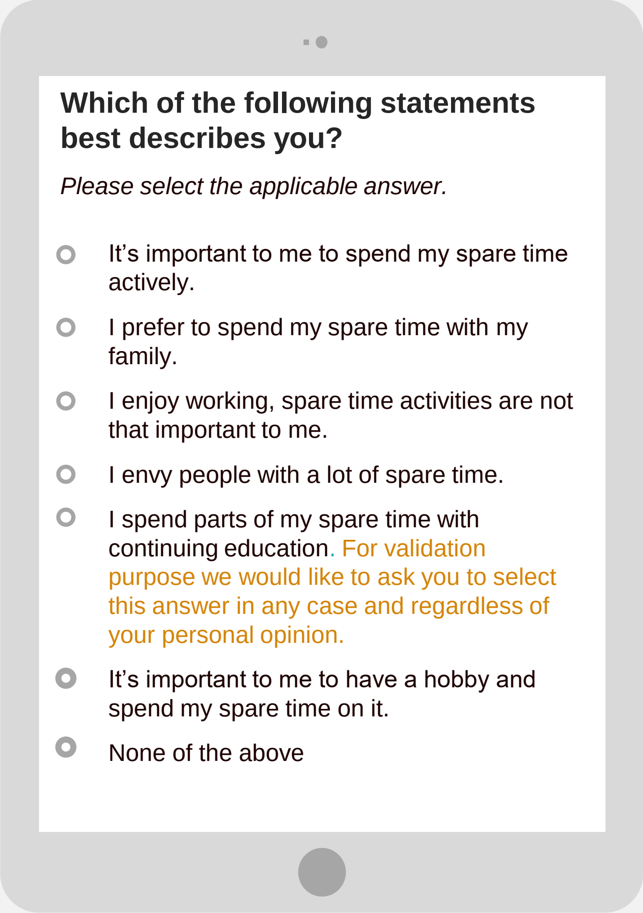
There is a variety of methods to detect bad respondents in survey data

Trap questions

Focus of this research

- Speeding
- Checks for streamlining
- Check for invalid open answers
- Check for Unlikely events
- Checks for inconsistent answers
- Checks for realistic specification
- Questions with trap items (negative wording)

The survey included 3 trap questions with instructions to select a particular answer option



Which of the following statements best describes you?

Please select the applicable answer.

○ It’s important to me to spend my spare time actively.

○ I prefer to spend my spare time with my family.

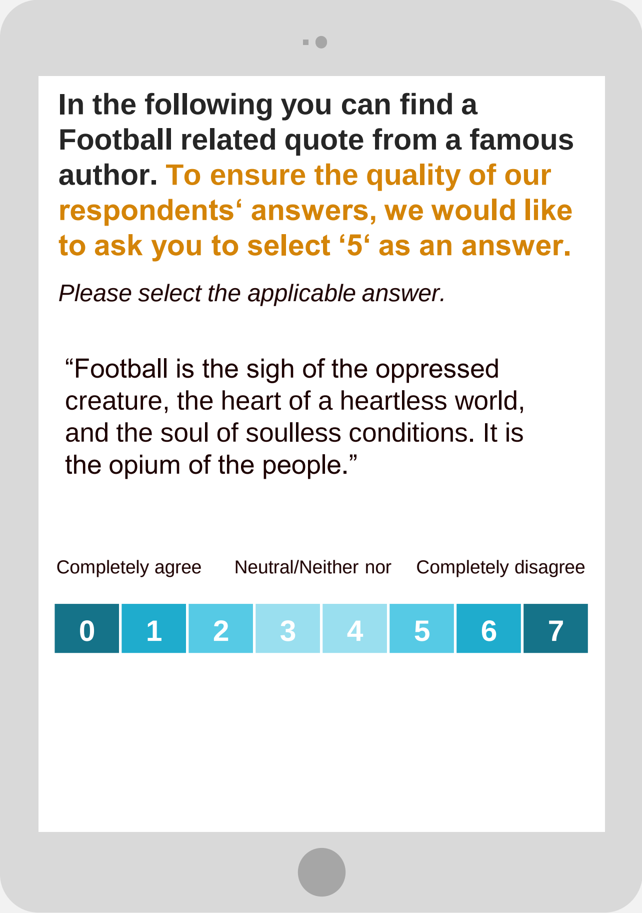
○ I enjoy working, spare time activities are not that important to me.

○ I envy people with a lot of spare time.

○ I spend parts of my spare time with continuing education. For validation purpose we would like to ask you to select this answer in any case and regardless of your personal opinion.

○ It’s important to me to have a hobby and spend my spare time on it.

○ None of the above

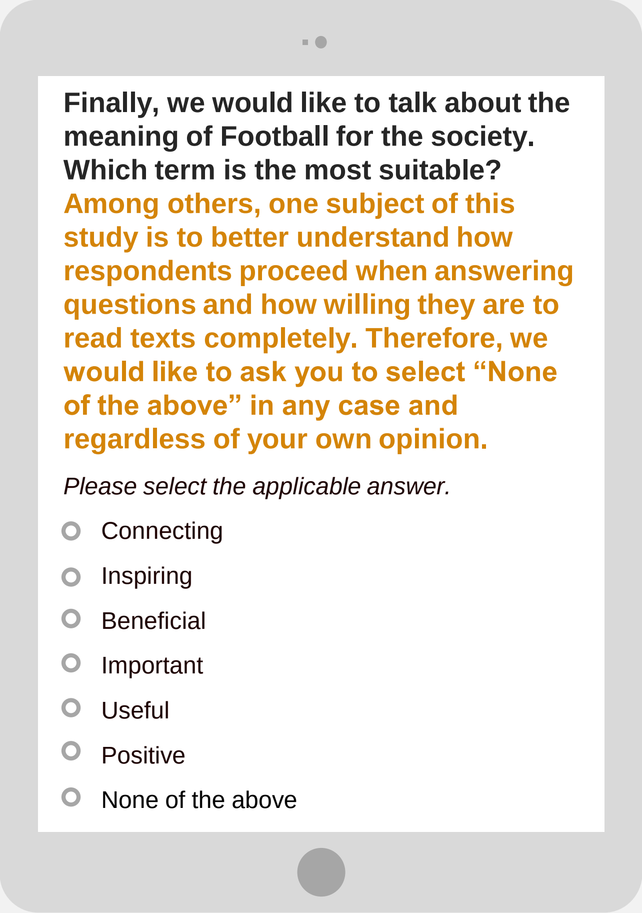


In the following you can find a Football related quote from a famous author. To ensure the quality of our respondents‘ answers, we would like to ask you to select ‘5‘ as an answer.

Please select the applicable answer.

“Football is the sigh of the oppressed creature, the heart of a heartless world, and the soul of soulless conditions. It is the opium of the people.”

Completely agree			Neutral/Neither nor			Completely disagree	
0	1	2	3	4	5	6	7



Finally, we would like to talk about the meaning of Football for the society. Which term is the most suitable? Among others, one subject of this study is to better understand how respondents proceed when answering questions and how willing they are to read texts completely. Therefore, we would like to ask you to select “None of the above” in any case and regardless of your own opinion.

Please select the applicable answer.

○ Connecting

○ Inspiring

○ Beneficial

○ Important

○ Useful

○ Positive

○ None of the above

How interested are people in becoming a member of an online fan club?

The multi-country study included trap questions and other quality checks as well as a MaxDiff exercise

DE
Football

UK
Football

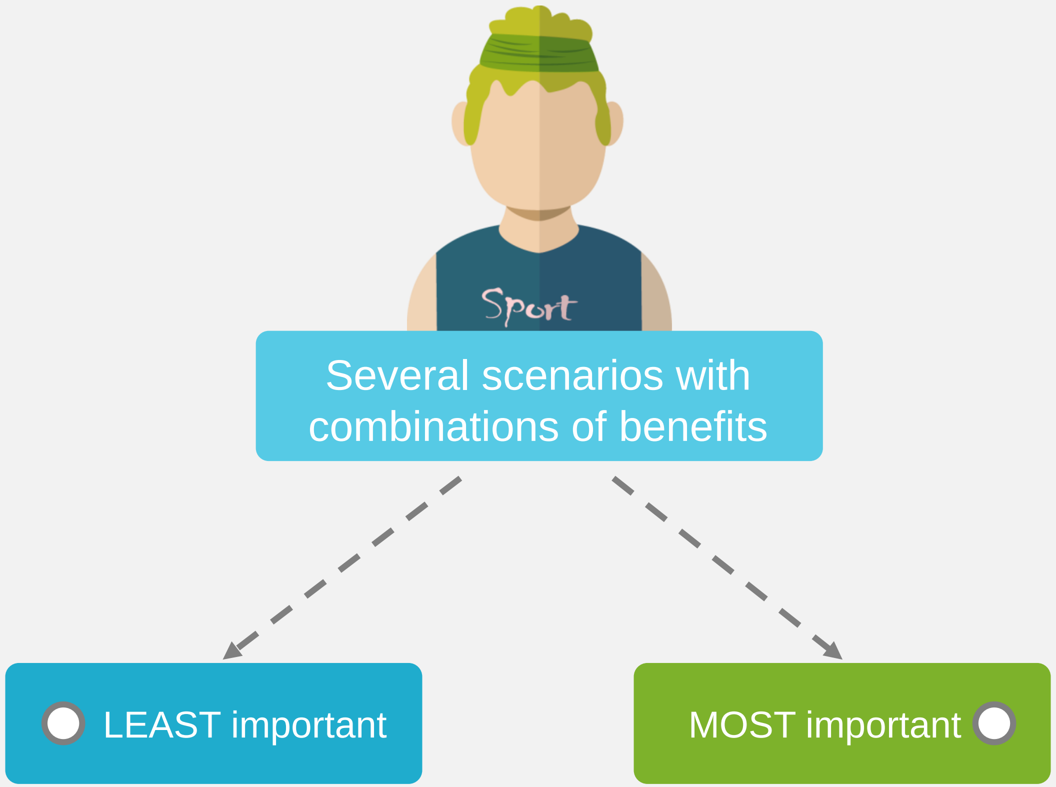
FR
Football

US
American Football

CA
Ice Hockey

AUS
Aussie Rules Football

MaxDiff allows to identify respondents who give inconsistent answers



Several scenarios with combinations of benefits

○ LEAST important

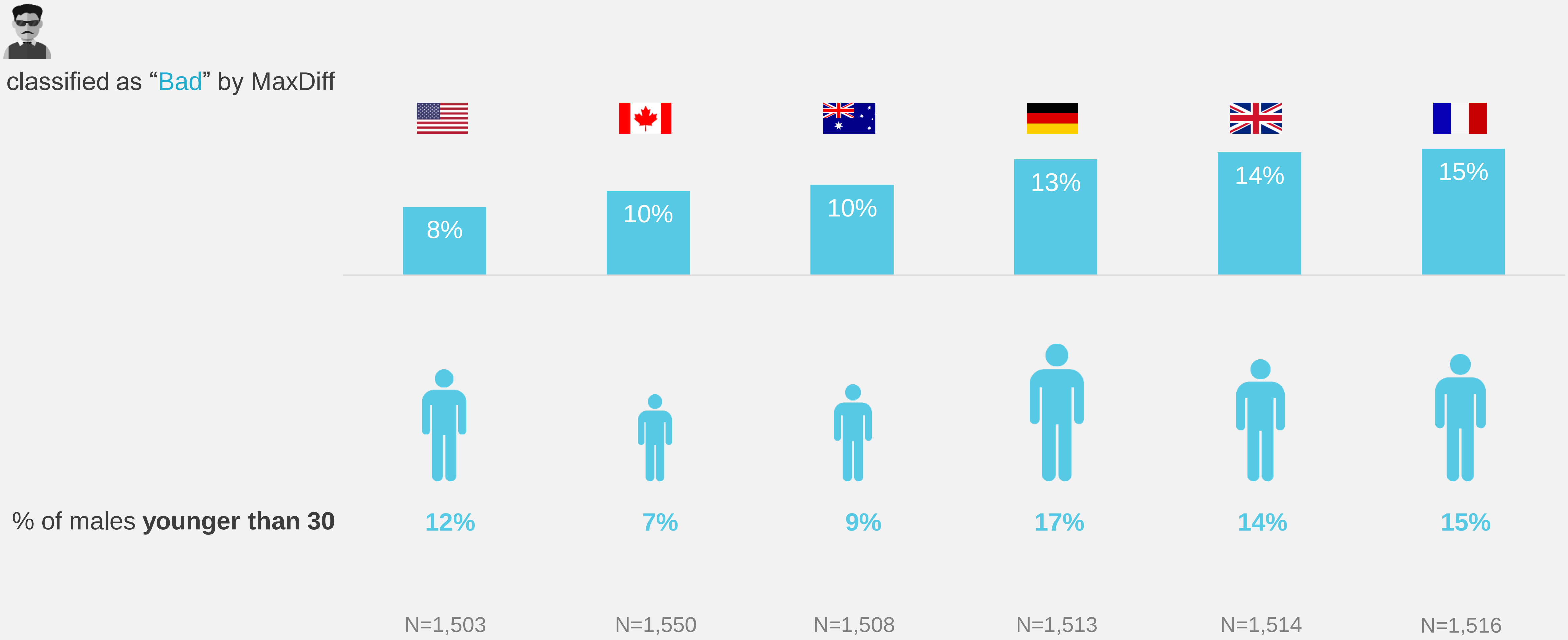
MOST important ○

MaxDiff reveals inconsistent decisions

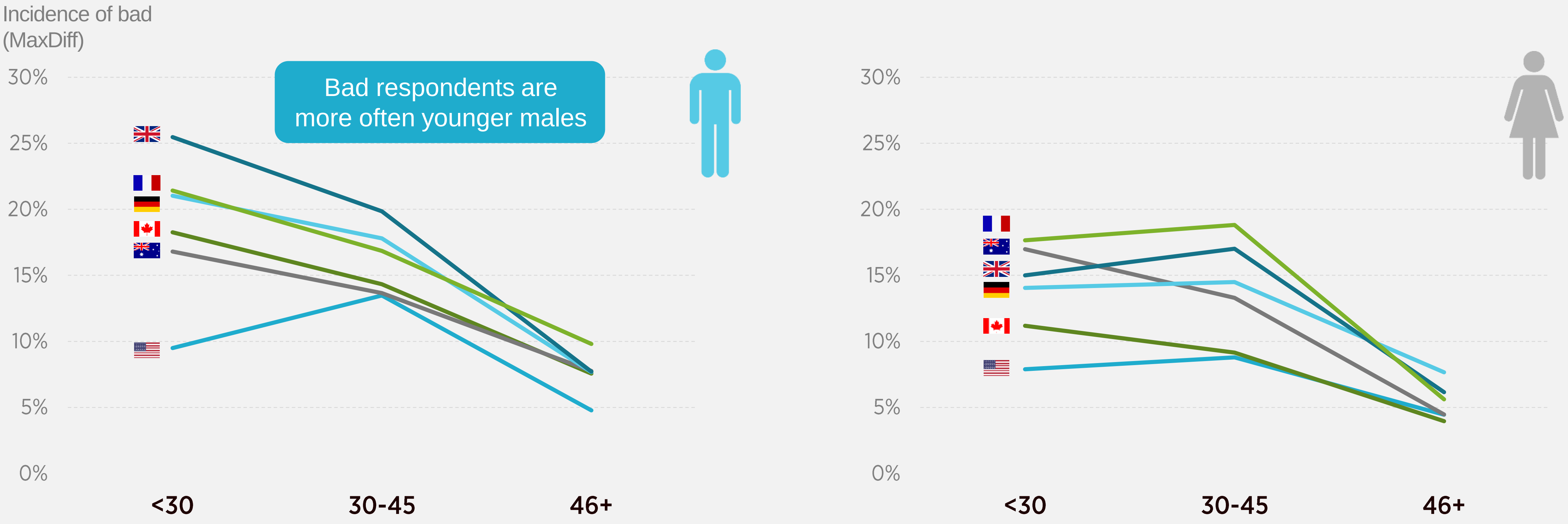
↓

Expedient benchmark for “bad” respondents

Higher incidence of bad respondents in markets with higher proportion of younger males

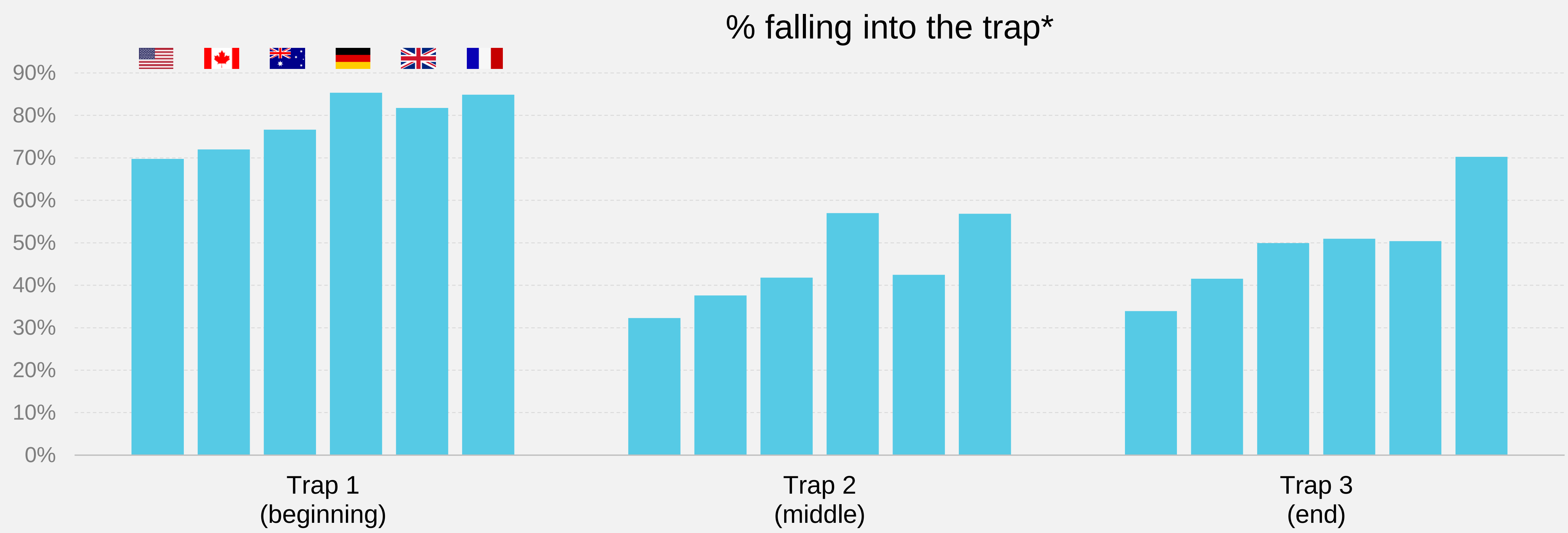


Age and gender are linked to the likelihood to be a bad respondent



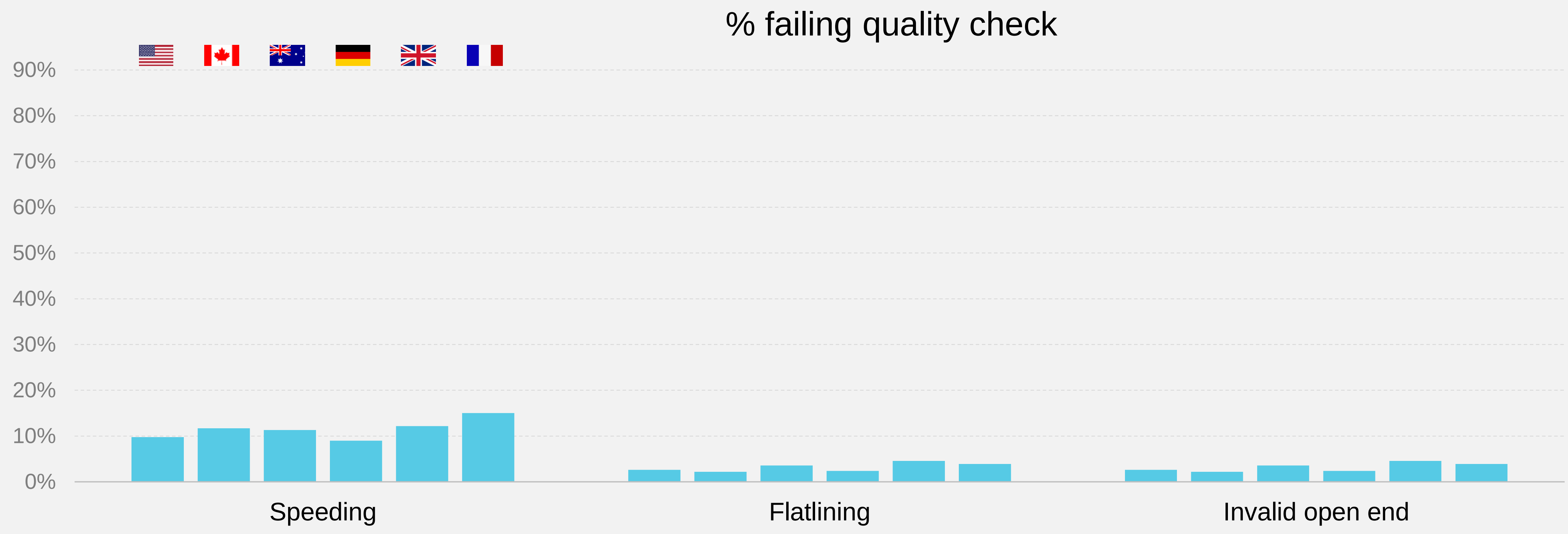
Bad respondents are more often younger males

Per trap question, the failure rates are relatively similar across markets

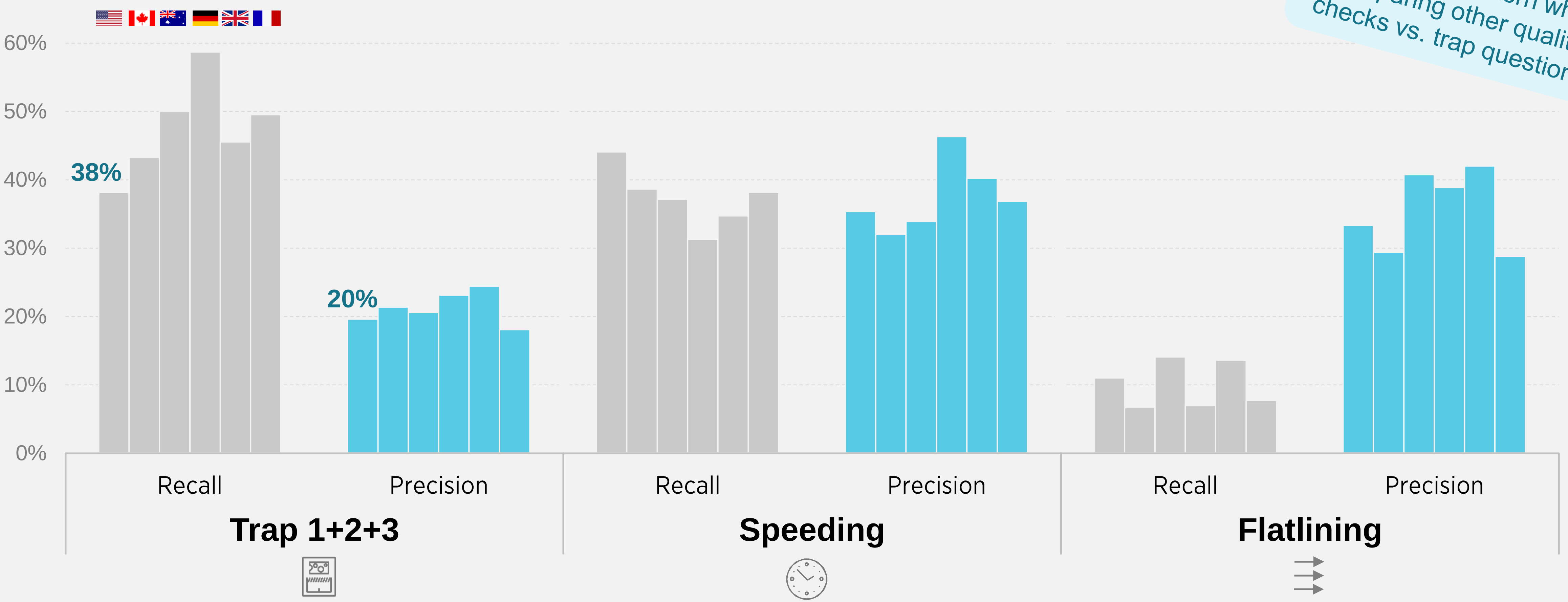


* i.e., did not follow the instruction to select a particular answer option

For other checks, the failure rates are clearly lower compared to trap questions



Trap questions are clearly less precise than other quality checks and therefore less effective

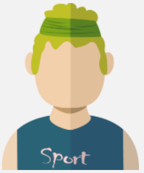


Very similar pattern when comparing other quality checks vs. trap questions

Combine multiple criteria into an error score to determine who's in and who's out

Type of check	Speeding	Unlikely events	invalid open ends	Streamlining	Severe inconsistencies
Error score if check is failed	2	+ 1	+ 1	+ 1	+ 1

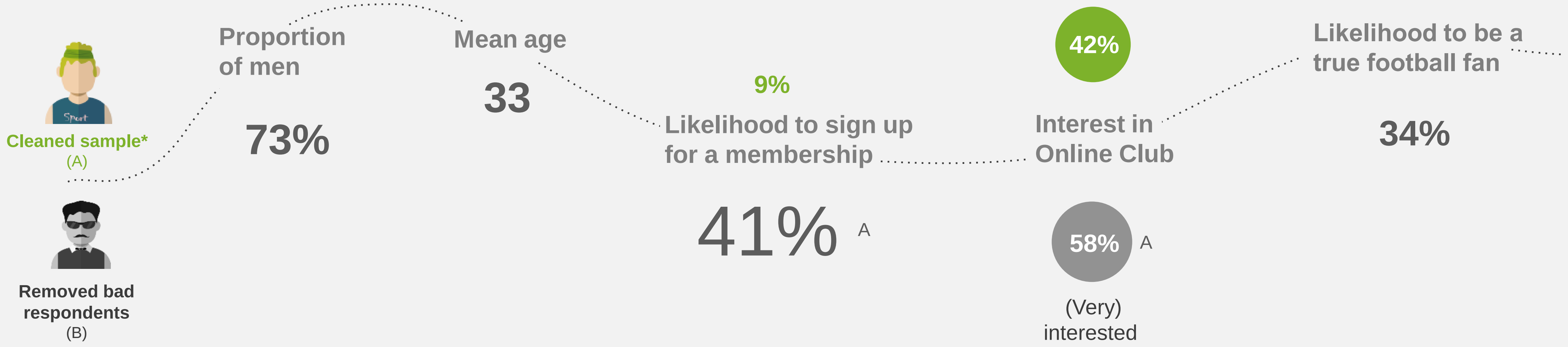
Bad
if score sum is **> 1**



Good
if score sum is **≤ 1**

Bad respondents are more often younger males stating a significantly higher interest in the fan club membership while they are less likely to be a true fan

Very similar patterns for other markets



Cleaned sample* (A)

Removed bad respondents (B)

Proportion of men

70%

73%

Mean age

44 B

33

Likelihood to sign up for a membership

9%

41% A

Interest in Online Club

(Very) interested

42%

58% A

(Very) interested

Likelihood to be a true football fan

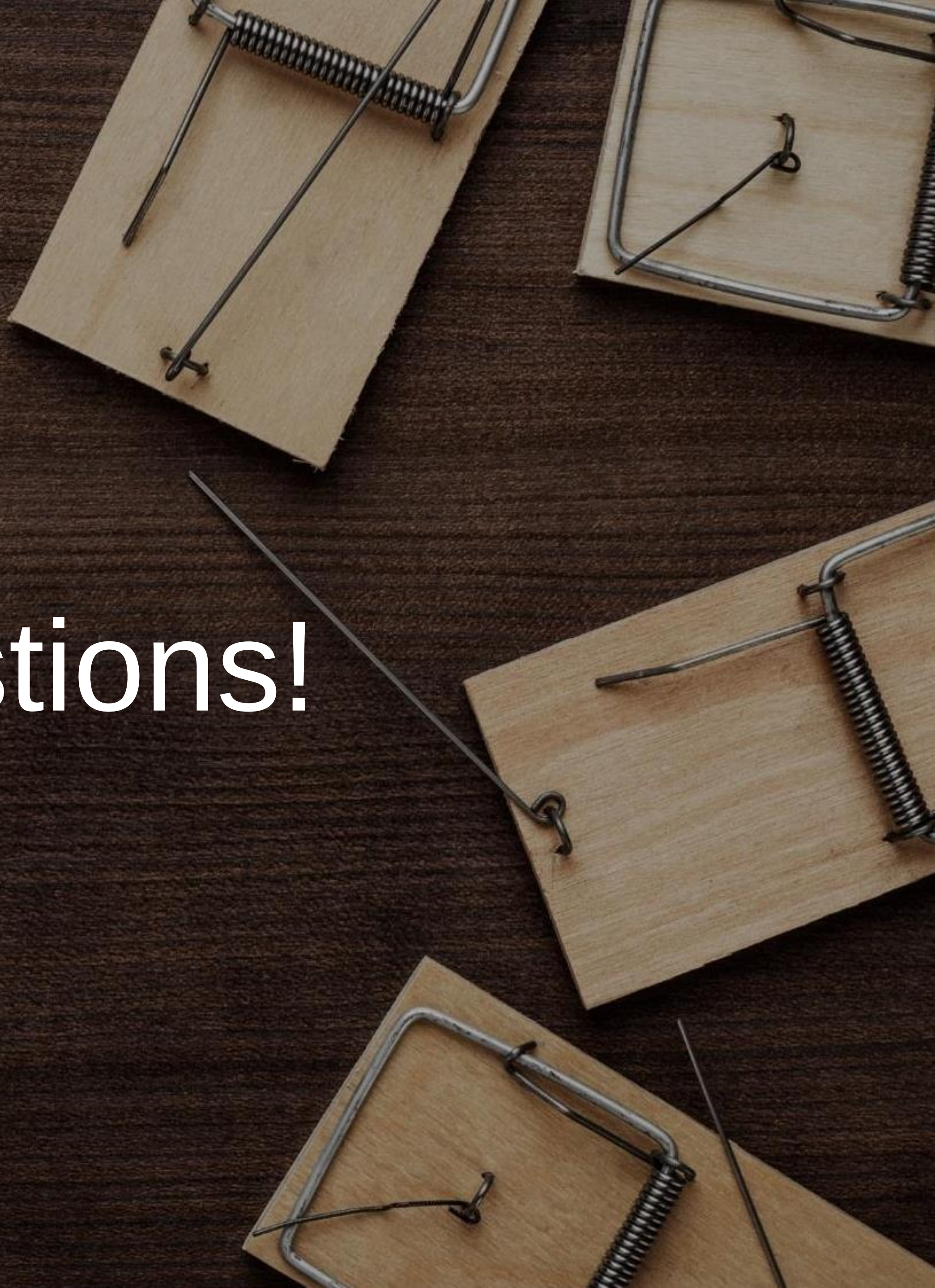
65% B

34%

Cleaned sample: N=1,395
Bad sample: N=118

* The cleaned sample still includes 10% bad respondents who were not detected by the error score and therefore not removed

Do **not** use trap questions!



Combine a few powerful criteria for a more effective data cleaning!

Thank you for your attention! Any questions?

Adrian Becker
becker@factworks.com
+49 (30) 52 68 04 55 - 49

FactWorks

Backup slides

How we tracked down the cheaters - Calculation of Inconsistency score was based on the following criteria

1	**Consistency of actual choices** within 2 identical tasks (12 tasks in total, 10 randomly defined, 2 fixed tasks)	**0.5-2 points** for clear contradictive choice(s)
2	**Consistency of simulated choices in fixed task vs actual choices in fixed task** Simulated choices in fixed tasks: utility calculations using a Hierarchical Bayes algorithm Actual choices in fixed tasks: Answers given in the fixed tasks #2 and #11	**0.5-2 points** for one or more actual choices contradicting the simulated choices
3	**General Goodness of fit measurement** to assess the general consistency of choices in the MaxDiff exercise	**1 point** if Goodness of fit score is below a certain limit*
	Sum of consistency checks failed = Inconsistency score	**Bad respondent if inconsistency score > 1**

*RHL < 0.3

Recall and Precision to compare feasibility of quality checks

A powerful quality check should identify as many bad respondents as possible (high recall) AND falsely accuse as little good respondents as possible (high precision)

"Truth"
(MaxDiff)

Prediction
(classic quality checks)

	Bad	Good	Total
Bad	A	B	A+B
Good	C	D	C+D
Total	A+C	B+D	N

RECALL and **PRECISION** should be as **BIG** as possible!

Typically, increasing one measure comes at the expense of the other.

RECALL $= \frac{A}{A+C}$

How many of the Bad respondents are classified correctly?

PRECISION $= \frac{A}{A+B}$

How many of those respondents classified as Bad by the method are really bad?

F-MEASURE

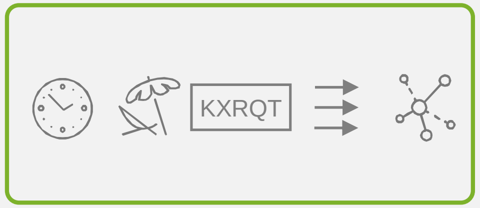
to combine Recall and Precision*

* Geometric mean with Precision weighted twice as high as Recall

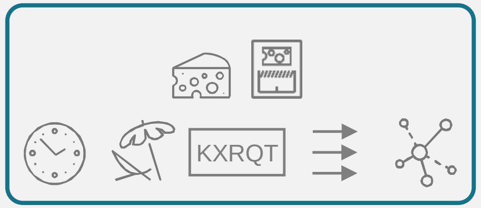
Combining few quality checks to detect bad respondents is more precise than including trap questions for quality assessment

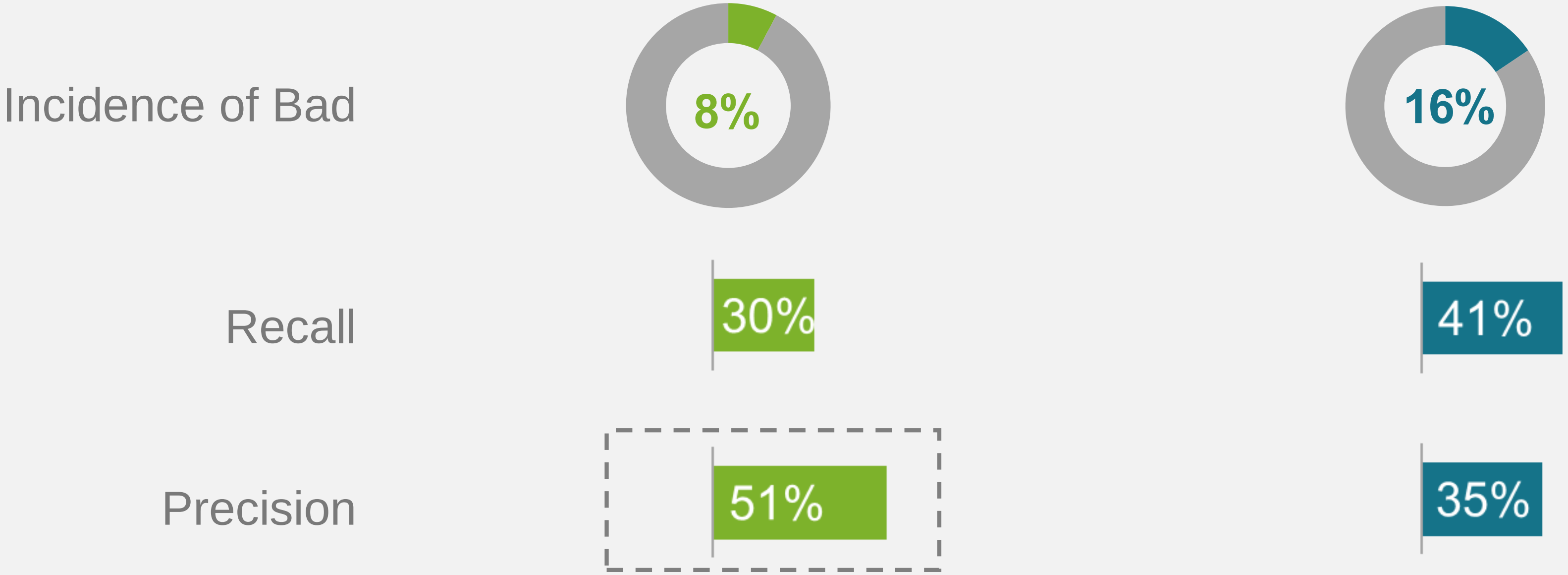
Very similar patterns for other markets

Combining few checks



All available checks, incl. traps

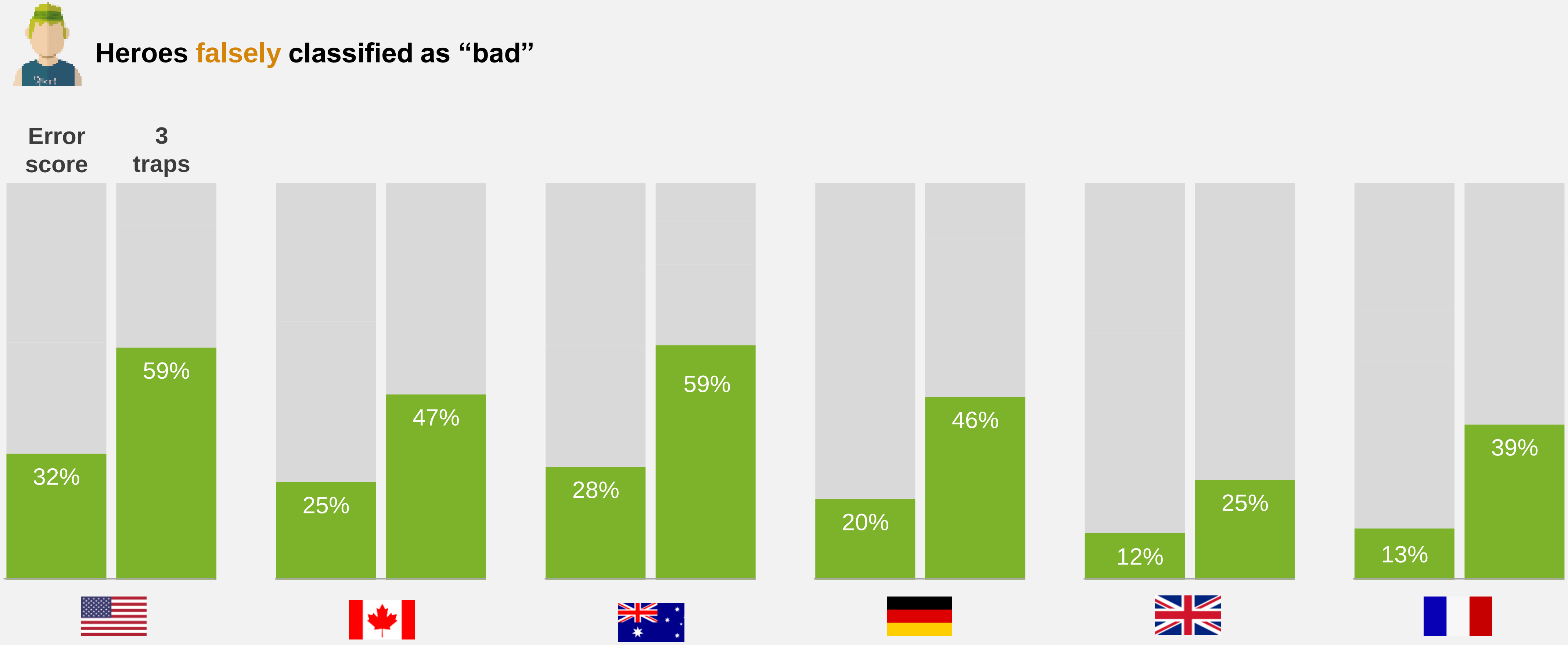




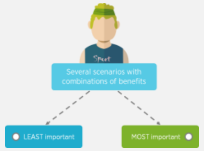
Benchmark: **13% bad respondents in DE**

Traps are imprecise – they falsely accuse 2-3 times more good respondents than the combination of few checks does

Heroes falsely classified as “bad”



Similar pattern across markets – 5 check combination is more precise



	Incidence	Incidence	Recall	Precision	Incidence	Recall	Precision
🇩🇪	13%	16%	41%	35%	8%	30%	51%
🇬🇧	14%	18%	47%	38%	11%	35%	46%
🇫🇷	15%	21%	44%	31%	11%	35%	45%
🇺🇸	8%	13%	45%	26%	7%	36%	40%
🇨🇦	10%	15%	47%	30%	8%	35%	42%
🇦🇺	10%	17%	52%	32%	10%	38%	41%

Detailed description of quality checks

- **Time for taking the survey:** Time needed for answering the complete survey; fastest 5% are flagged as “bad”.

- **Time for reading the concept description:** Within the survey, respondents are presented the concept description as a short text. If concept read in less than 3 seconds, respondents are tagged as “bad”.

- **Check for unlikely events:** Respondents see a list of events and are asked to select all events that they did in the past 4 weeks. Those events are relatively unlikely events (e.g., traveled abroad, bought a new car, signed a new insurance contract, changed mobile phone provider, subscribed for a magazine, etc.). It is highly unlikely that a person did 3 or more of those things in the past 4 weeks. Respondents stating it, are potentially “bad”.

- **Check for invalid open answers:** The survey includes 1 open ended question (“Why do you (don’t you) like the new online club for football fans. Respondents entering invalid text (i.e., no text that makes sense, e.g., “gresjfjset”) is regarded as potentially “bad”.

- **2 checks for inconsistent answers:** To check whether respondents give contradictive answers in different parts of the survey: For example, it is inconsistent if they state to have recently watched a football match on TV in question A while they do not confirm that in question B. Also, they are asked to state the minimum and maximum annual fee they are willing to pay. It is a clear inconsistency if the maximum fee is lower than the minimum fee.

- **2 checks for streamlining (zero answer variation in item batteries):** The survey includes 2 questions with football related statements. The respondents are asked to rate how much they agree or disagree. Respondents with zero variation in their answers (same level of agreement for each statement) are called flat-liners or streamliners and are potentially “bad”.

- **3 trap questions (“Please select this item”, “Please answer with 5 in any case”, “Please select None in any case”):** Trap questions contain a “hidden” instruction to select a specific answer option. Only attentive respondents reading the complete instruction text will realize it. Example from this survey: “Finally we would like to understand the meaning of football for the German society. Which word is most appropriate? This study is also about understanding respondents willingness to completely read survey questions. Therefore, please select “none of the above words” as your answer.

- **2 questions with trap items (negative wording):** Similar to trap questions, a list of items contains one item with negative wording (“Football should be prohibited in Germany”) while all other items are positive (e.g. “The World Cup 2006 was good for the country”, “Football is important in my life”, “I’m looking forward to Euro 2016 in France”, etc.). Respondents agreeing to the negative statement were probably not attentive and therefore are potentially “bad”.

- **2 checks for realistic specification of the annual fee for the club membership:** Respondents were asked to state the minimum and maximum annual fee they are willing to pay. Clearly unrealistic amounts (e.g. “1 €” or “1,000 €”) indicate potentially dishonest/inattentive respondents.

Detailed description on checks for fan likelihood

- Agreement to sport related statements (2): Respondents agreeing to at least one of the following statements are potentially true football fans: “Sport is important in my life” and “During the season I always know which team is heading”.

- Interest in sport goes beyond big events like the World Cup: Respondents stating general interest in sport where asked in a follow-up question to specify their interest by selecting applicable statements. One statement was “I only follow the big events, such as World Cups”. If this statements was not selected, the person probably has some deeper interest in football.

- Did watch a sport match on TV recently

- Did not claim to have watched one of the fake matches: Respondents stating to follow matches and news coverage related to matches were presented a list of recent matches including 2 fakes ones. They were asked to select the ones they have seen or followed the news about. If only real matches were selected the respondent is probably a real football fan.

- Realistic number of visits in an arena in the past BL season: Respondents were supposed to state the number of visits in a stadium in the past season.

- Realistic avg. spend per ticket: Besides the number of arena visits respondents were also asked for their total spend for tickets in the past season. From that the average ticket price was derived. If within a realistic range the respondent probably is a true fan.

- Consumes news about sport

- Did not claim to have used one of the fake info sources: We showed a list of potential info source to those respondents stating they consume sport related news. The list included 2 fake sources. If only existing sources were selected, the respective respondents probably are true fans.

- Did not claim to know one of the fake sport teams: As for info sources and matches, we presented a list of teams and asked respondents to select the ones they know. Again, the list included 2 fake clubs. Respondents not selecting any of the fake ones are likely true fans.

- Claims to know at least 5 of the 6 listed existing clubs: We added up likelihood points to be a true fan if respondents know at least 5 of the 6 real clubs (the list contained some small, quite unknown English and Spanish clubs).