Sprekend Nederland

a recent multi-purpose collection of Dutch speech

Frans Hinskens, Meertens Instituut, Vrije Universiteit
David van Leeuwen, NovoLanguage, Radboud University Nijmegen

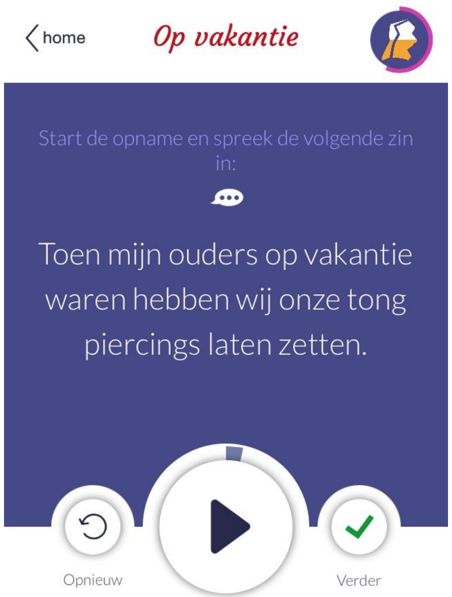
What is *Sprekend Nederland?*

- A collection of speech recordings, speaker metadata and perception/attitude questionnaires collected in 2016
  - crowd sourced
  - approximately 10 000 participants
  - all in Dutch / mostly in The Netherlands
- A project at the Dutch broadcast organisation NTR,
  - aiming at registering all spoken accent in The Netherlands
  - hoping to debunk prejudisms against stereotypical regional accents
  - resulting in various productions on social media and national radio and TV
- A co-operation between scholars from various disciplines
  - linguistics, phonetics, sociolinguistics, social psychology, sociology, speech technology
  - no funding

Basic idea

Everybody in NL downloads and runs a free *app*, which

- guides participants to a sequence of interactions, including
  - giving consent to use data for research and development
  - recording an utterance (reading a prompt text / naming a picture / making a description)
  - providing some personal data (age, sex, origin, social attitudes)
  - listening to an utterance, and judging the other speaker on linguistic and sociological aspects
- should somehow be *fun*, by
  - obtaining other participants' (filtered) feedback about one's own accent
  - including various language games ((tongue twisters, riddles, jokes etc.)
- could be run in multiple sessions over longer time
  - content naturally organized in different themes
  - dynamic functionality and content



Start de opname en spreek de volgende zin in:

Toen mijn ouders op vakantie waren hebben wij onze tong piercings laten zetten.

The partners and their tasks

- broadcast organisation NTR
  - initiation, media production, sponsor for app development and operation
- academia
  - inventory of research questions
  - experimental design
  - stimulus material
  - progress monitoring
- app-production company *Alledaags*
  - front end smartphone app: Android and iOS
  - back and servers: distribution of tasks and database storage of audio and responses
- archive Sound and Vision (Nederlands Instituut voor Beeld en Geluid)
  - long term storage and access to the data

Aims of *Sprekend Nederland*

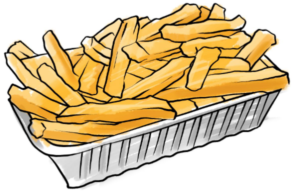
1. Assembling a huge and rich database for scholarly research, containing
    - spoken modern standard Dutch from as many different speakers as possible – in as many different (geographical, social, stylistic, and/or ethnic) varieties as possible – in Haklay's (2013) 4-level model: 'citizens as sensors'
    - the perception of and attitude towards all these varieties - in Haklay's (2013) 4-level model: 'citizens as interpreters'
2. Informing a general audience about geographical, social, stylistic and ethnic variation in spoken modern standard Dutch. Some first findings have eventually been communicated to the larger audience (the 'crowd') via *Kennis van Nu* TV show, Facebook posts and related social media

Approach

- As part of the preparation, we sent out questionnaire to researchers, directors of research and deans in NL in wide range of disciplines
  - what kind of research question do you have that could be researched using SN data?
  - what interactions between app and participant would this require?
  - what kind of stimulus material would you need?
  - what meta-data do you need to have about the participants?
- Prioritized interaction-types in app
  - Each type requires code to be implemented in app, limited resources
  - Answer types:
    - yes/no
    - multiple options
    - 7-point Likert scale
    - number
    - location on map (pan/zoom)

Approach (continued)

- Decided on stimulus-data in app
  - Stimuli:
    - 10 sentences plus a set of 44 words, covering 5 major instances of supra-regional phonemic variation,
    - 122 loan words
    - 278 words covering all consonant-vowel combinations occurring in Dutch
    - 2071 sentences for lexical variability
      - originally pool of 48 million sentences requested
    - 130 pictures to be named for eliciting regional lexical items
    - 9 assignments to describe something, for eliciting spontaneous speech

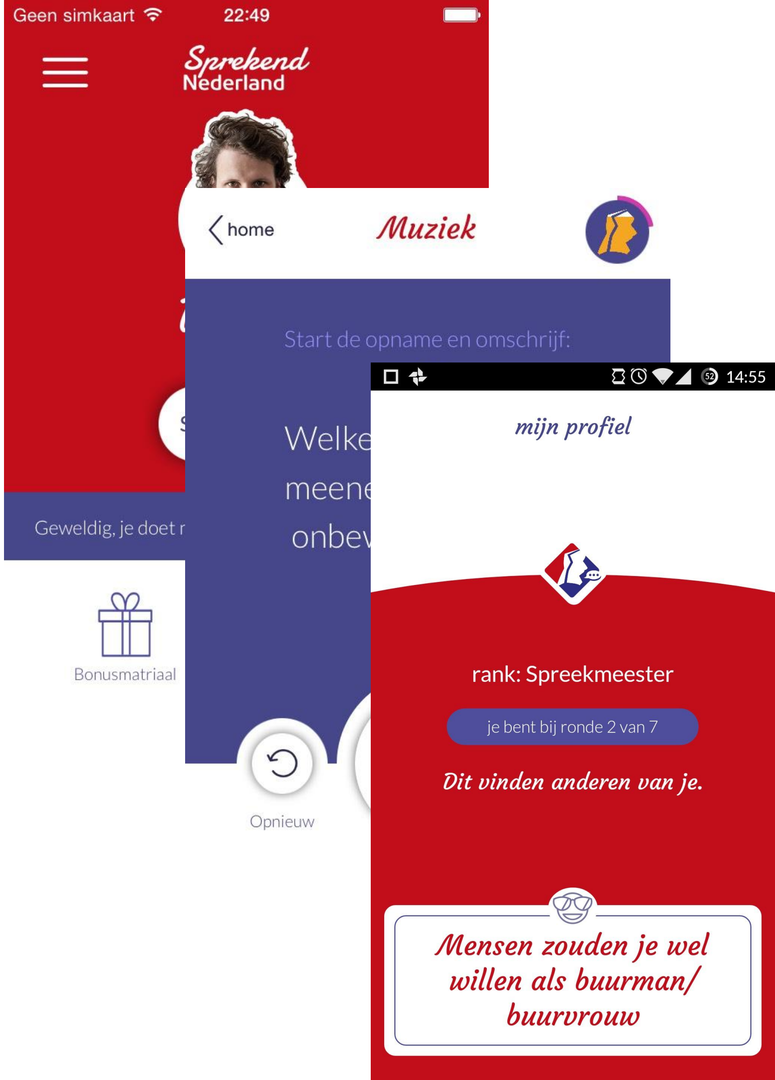


App design: different interests

- NTR
  - fun to use
  - themed structure
  - “sell” well on TV / radio / internet
- Researchers
  - all speakers record all regional variation sentences and words
  - all speakers name all pictures
  - all speakers record all loan words
  - all speakers answer all sociological attitude questions
  - as many speakers record many unique sentences
  - all speakers judge all other speakers on all attitude aspects for all speaking styles
- App production
  - as few as possible user-interface elements
  - no complicated run-time server decisions

Consensus strategy

- NTR negotiates and decides
  - NTR - Researchers, prioritize and select
    - metadata questions
    - stimulus material
    - attitude and perception questions
    - sociological attitude questions
    - speaker - listener distribution
  - NTR - App production, using SCRUM methodology
    - interaction flow
    - theming, styling, feedback, gamification
    - question / stimulus order
    - server operation decision at production time
  - NTR
    - rest



Start de opname en omschrijf:

Welke meene onbev

Geweldig, je doet r

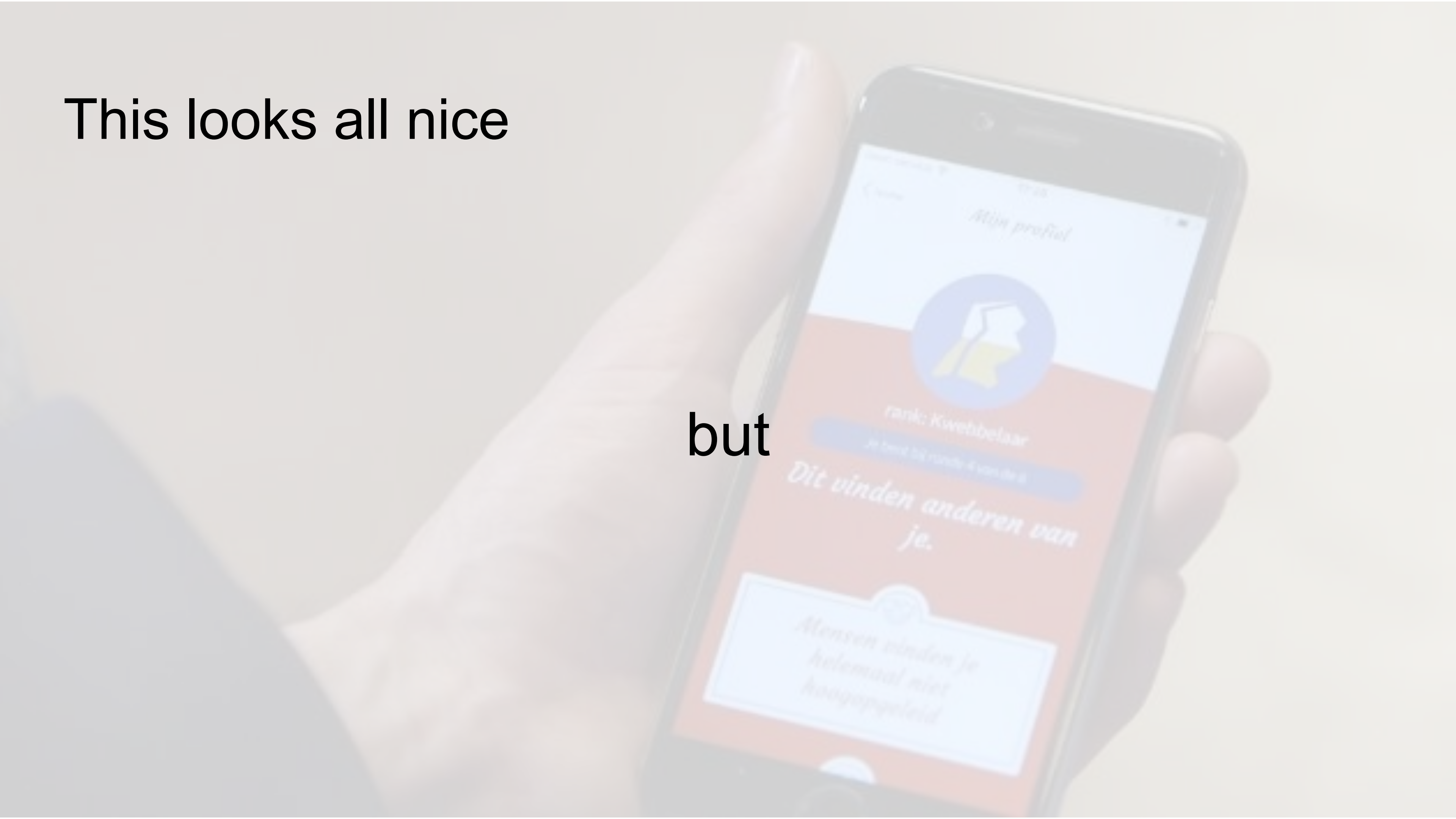
Dit vinden anderen van je.

Mensen zouden je wel willen als buurman/ buurvrouw

Some numbers (final, including under-18)

- 1 dec 2015 -- 31 dec 2016
  - Approximately 5 nation-wide media events
- 17 885 participants registered
  - 10 025 participants made at least 1 recording (56%)
  - 12 979 participants gave at least 1 answer to a question (73%)
- 292 863 recordings were made (average 29.2 per recording participant)
  - 528 hours of audio, average 6.5 sec per recording
- 1 744 588 answers to questions in the app were given
  - 9% to personal questions (age, sex, origin, attitude), average 12.3 per answering participant
  - 89% to attitude questions about other speakers

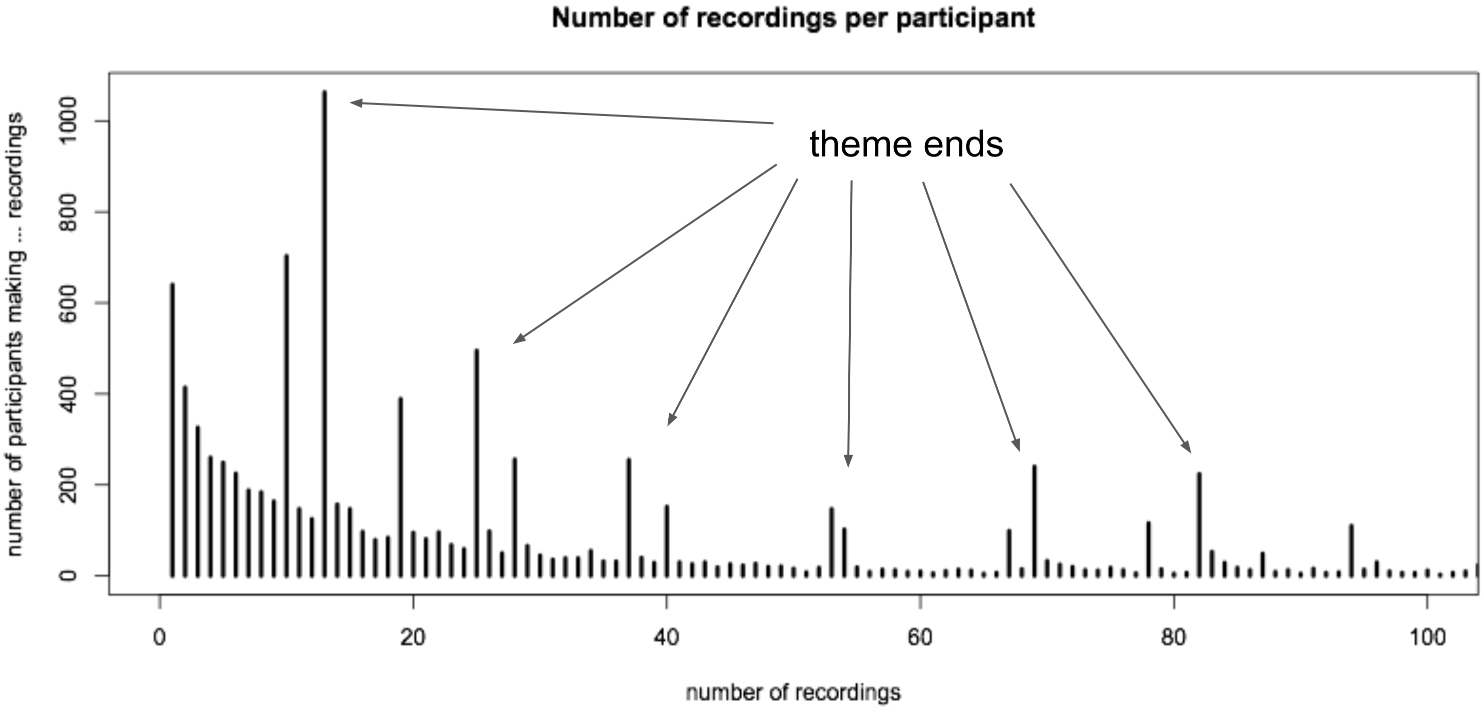
This looks all nice


Mijn profiel
rank: Kwebbelaar
Dit vinden anderen van je.
Mensen vinden je helemaal niet hoogopgeleid

but

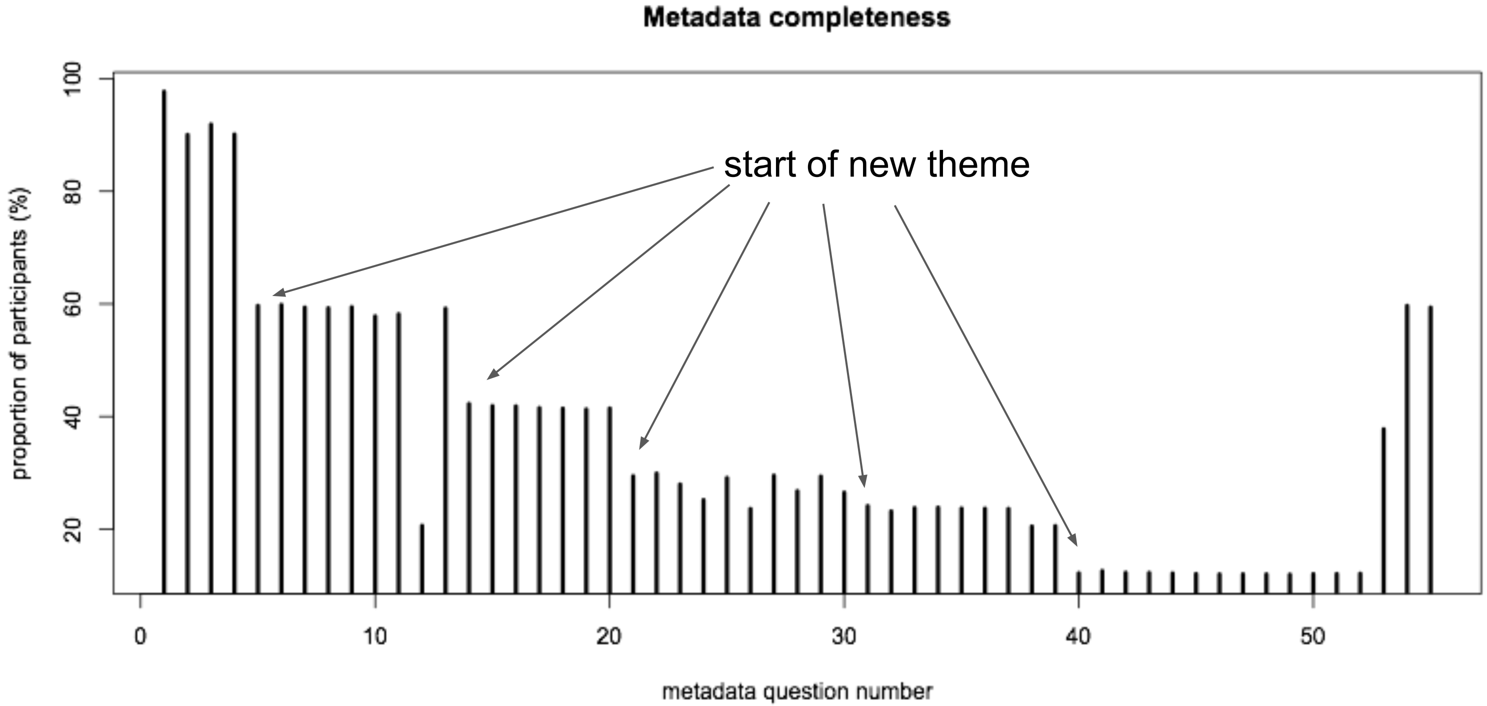
Participants were free to quit at any time

- a not unreasonable condition in IRB-approved research involving subjects



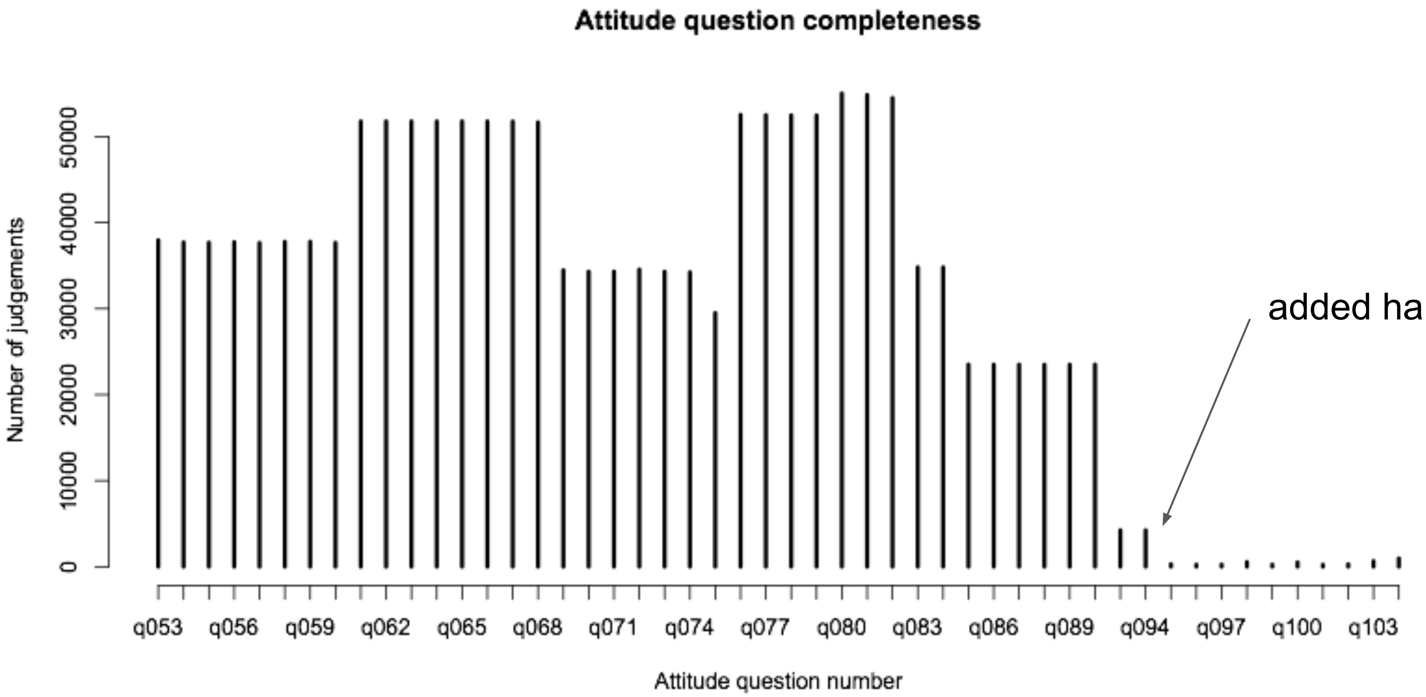
Interactions were in same order for all participants

- as per themed design of app



The feedback-to-implementation time (vv) was long

- No point in blaming any specific partner



added halfway the project

The Sprekend Nederland approach: pros

- Research data virtually without proposal / rebuttal / costs
- Largish sample of the population
  - slightly different from white / male / 20-year old / psychology student (WEIRD)
- Leverage wide distribution of high-quality data acquisition devices
  - i.e., smartphones
- Large influence to decisions about
  - experimental design
  - stimulus material
  - questionnaire data
- Research gets attention in traditional media
  - wide layperson audience
- Generally fun to do
  - not in the standard research infrastructure

The Sprekend Nederland approach: cons

- Preparations have not always received the usual academic scrutiny
  - broadcast organisations have production deadlines
  - … but lose interest after broadcast has taken place
- No complete control over
  - implementation
  - recruitment of subjects
  - completeness (sufficient socio-biographical metadata for some 3500 participants, too few socio-biographical metadata for some 7000 participants)
- Resulting in skewed databases
  - providing interesting challenges to statistical analysis
- Hardly any human quality control / annotation
- Data not owned by research institution
  - different guarantees concerning data persistence and quality
  - no clear path towards ethics approval

Conclusions

- Participation in such a project was *fun*
    - at least, for us researchers
- A large volume of data can be collected in a short time for little money
    - but distributions are skewed
    - many NAs in (meta)data
- Disclosing the data is quite an effort
    - structured, but complex, relational database
    - acquired a small NWO KIEM subsidy to prototype a faceted data browser and explore the data some more
- Advice for similar endeavors
    - Be very careful with (anonymous) feedback. People are harsh, judge stereotypically, and this is probably not an incentive for participating
    - keep a close eye on technical development and the data distribution as it comes in
    - negotiate a strong position in experimental design decisions