Paradata: Measurement

Fundamentals of Data Collection II

January 13, 2025

Frederick Conrad

Introduction

- The measurement process generates *respondent paradata* (e.g., Couper 2008) that can help identify *measurement error*
    - i.e., discrepancy between what is observed and the true value
- Examples: Response times, respondent speech (interviews), respondent mouse movements (online, self-administered)
- Respondent paradata are typically indirect measures of quality but can alert researchers to potential problems with the data
- *R*-paradata often of unknown structure and size before collected, e.g., mouse clicks, but some are regular, e.g., response times

Outline

- Interviews
  - Response time (RT) can be affected by characteristics of items, respondents and interviewers
  - *R*’s speech can predict response difficulty (and response accuracy)
  - *R*’s visual behavior (gaze aversion) can predict reliability of responses
- Web surveys
  - *R*’s inactivity can indicate difficulty and trigger help
  - *R*’s mouse movement can indicate comprehension difficulty
  - *R*’s multitasking (leaving survey and returning) can be related to item nonresponse
  - *R’s* very fast answers can be discouraged with interactive prompting

Response Times (RTs) Can be Related to Measurement Error

- Short (fast) RTs are associated with Acquiescent Response Style (e.g., Basilli, 2003), lack of motivation to answer accurately (Yan & Tourangeau, 2008), inaccurate answers (Ehlen, et al., 2007)
  - But can reflect fluency and ease of responding (Olson & Parkhurst, 2013)
- Long (slow) RTs can indicate uncertainty, ambivalence, inaccurate answers, and are associated with item non-response such as DK (Draisma & Dijkstra, 2004; Olson & Parkhurst, 2013) and inaccurate answers (Ehlen, et al., 2007)
  - But can indicate thoughtful responses
- Meaning of RTs is less about whether fast or slow in absolute sense but whether are faster or slower than standard of comparison

Impact of Items, Respondents, and Interviewers on RT

- Couper & Kreuter (2013) analyzed RTs for each observation (i.e., each administration of each question to each respondent) in National Survey of Family Growth
  - Items (questions) nested within respondents nested within interviewers
  - Modeled RT on basis characteristics of *Item*s (e.g., fixed options, numeric, multiple response, open), *R*s (e.g., gender, age, education, race), and *Iwer*s (e.g., education, race, experience, Spanish speaker)
  - Source of data about characteristics: *Item*s from Blaise audit trail, *R*s from survey data, *Iwer*s from interviewer questionnaire
  - Modeled separately for female and male *R*s; all *Iwer*s female
- Lots of statistical power so usual significance levels may be misleading

Selected Results* for RT due to *Item*, *Respondent*, and *Iwer* Characteristics

- Items:
  - Fixed Choice, Numerical, and Multiple Responses < Open Response
  - Self (ACASI) administration < *Iwer* (CAPI) administration
- Respondents:
  - White < non-White
  - Never Married > other Marital status
  - more Educated < less Educated
  - Older > Younger
- Interviewers:
  - more Educated < less Educated
  - more Experience < less Experience

*All results significant at p < .001

Conclusions: RT as function of *Item*, *Respondent*, and *Iwer* Characteristics

- Many automatically derived *Item* characteristics vary with RT
  - But accounts for small proportion of RT variance
  - Measures of item sensitivity or task complexity require human judgment
- Results largely replicate results from more controlled studies – especially for *Respondents* – but data in current study collected in actual field (production) conditions
- *Iwers* contribute independently to completion times although demographics account for small fraction of variance
- May be possible to flag *items*, *R*s or *Iwer*s who are much slower or faster than expected to investigate further or intervene in real time

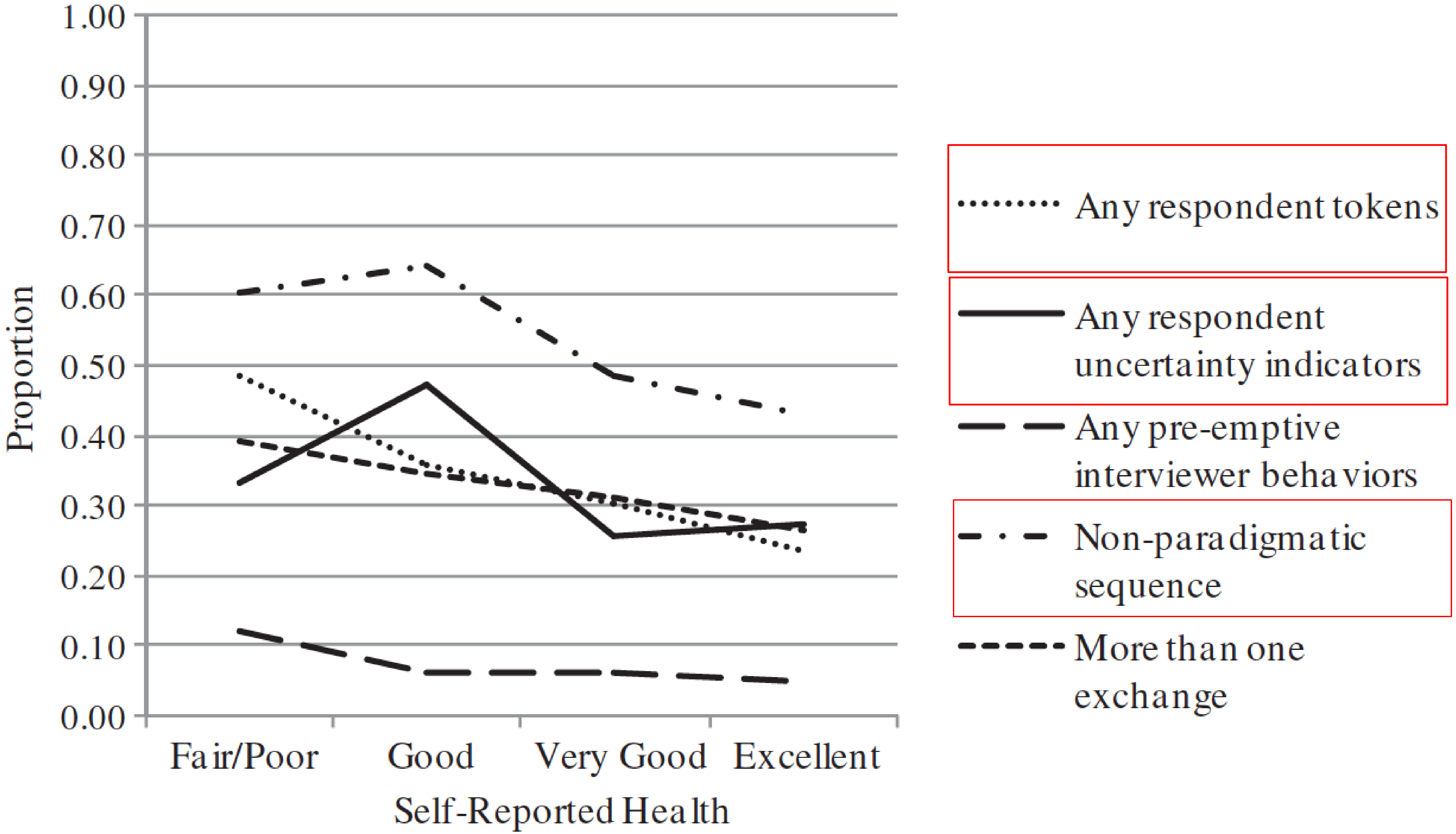
Respondents’ speech can predict and inform their answers

- *Rs*’ words and paralinguistic utterances (e.g., *um* and *uh*) can indicate how well understand *Q* or amount of difficulty answering
- Garbarski, Schaeffer, & Dykema (2011) proposed that valuable info about *Rs’* thinking may be found in *I-R* interaction preceding response; focus on Self Reported Health (SRH) question
  - “Would you say your health in general is excellent, very good, good, fair, or poor?”
  - Preceding interaction should be especially informative when difficult for *R*s to map their actual health to response categories
  - Should be easy to respond “excellent” but harder to choose between “good” and “very good”

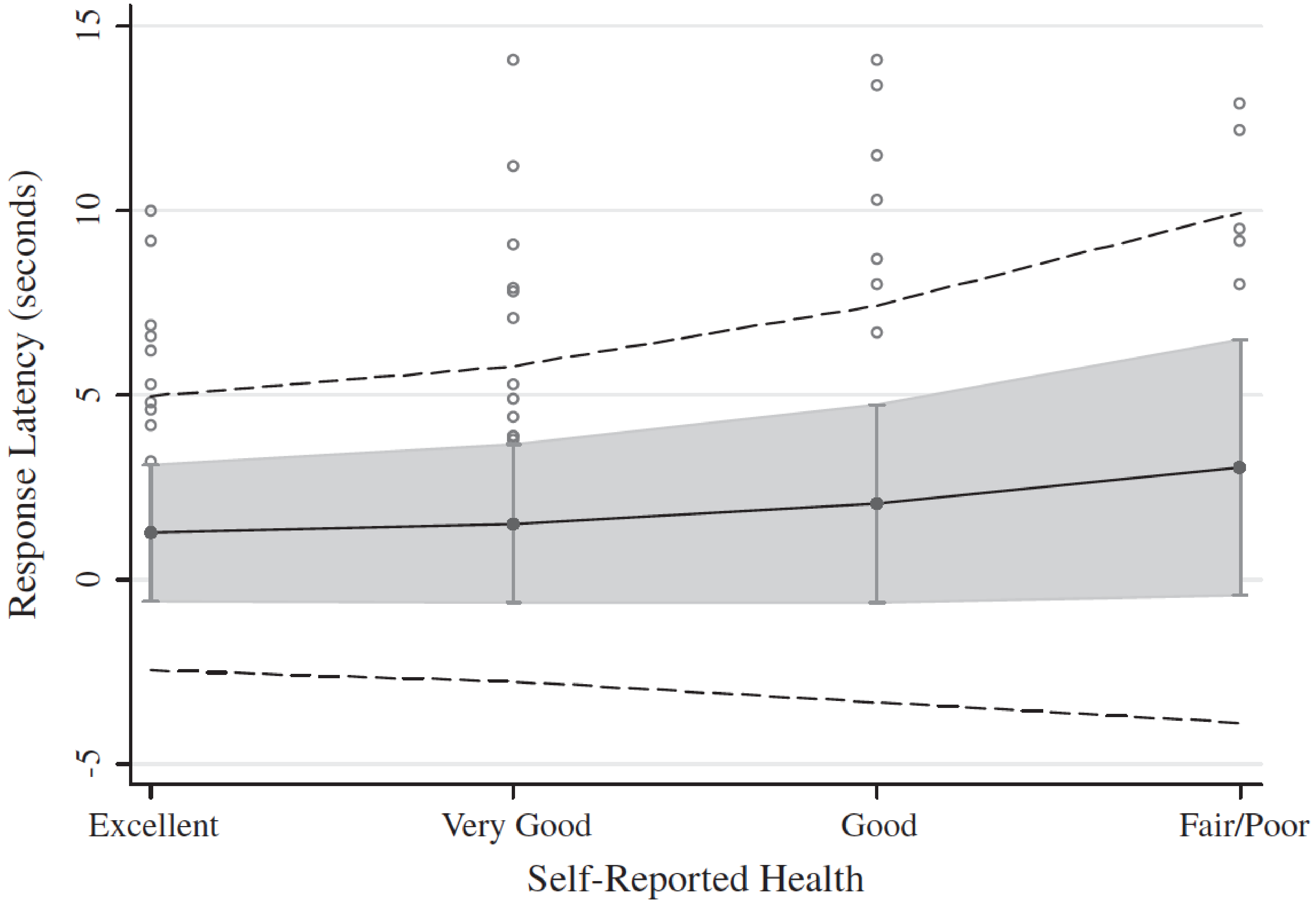
*R*s' speech can predict and inform their answers (2)

- Main prediction: when health circumstances more complicated, i.e., less than excellent, will observe more problematic speech and slower RTs
- Garbarski et al. (2011) operationalized "problematic speech" as
  - disfluencies (*um*, *uh*), indications of uncertainty (e.g., reports rather than answers), hypothetical responses (e.g., "pretty good"), ranges (e.g., "good to very good"), mitigating phrases ("I guess excellent")
- Analyzed problematic speech preceding SRH response among 355 phone *R*s from 2004 Wisc. Long. Survey

Proportion Responses Preceded by Problematic Speech



Mean (and SD) Response Latency by Each SRH Option



*R*s' Visual Behavior can Predict Reliability of Answers

- In-person interviews produce both *verbal* and *visual* paradata that seem to indicate response difficulty
- Visual paradata can include:
  - Facial expression
  - Head movement
  - Gaze aversion
    - e.g., Glenberg, Schroeder & Robinson (1998)
- Are verbal and visual paradata functionally similar? Redundant?
- If *R* produces visual paradata as evidence they need help, should be more frequent in Conversational Interview – where help is possible -- than Standardized Interview – where *Iwer* cannot clarify concepts

*R*s’ Visual Behavior Can Predict Reliability of Answers (2)

- Lab study (Schober et al., 2012) compares ability of
  1. *disfluencies* and gaze aversion in Conversational (CI) and Standardized (SI) interviews
  2. disfluencies in telephone and In-person interviews
- to predict reliability of initial answer within *Q-A* sequence
  - i.e., when *R*s’ speech includes two responses within sequence, are they the same?
- Design: 2 (Int. Tech’s: CI vs. SI) x 2 (Modes: Telephone vs. FTF);
  - questions re-administered (paper q’aire) after interview to evaluate *response change*
  - Paper q’aire Included definition of key question concepts

Results

Verbal

- *R*s produce more disfluencies (per 100 words) on phone (12.3) than FTF (6.4)
  - perhaps because no visual cues available to *R*s
- *R*s produce disfluencies in more answers in CI (55.4%) than SI (43.8%)
  - perhaps because they know conversational *Iwer*s can help based on cues
- *R*s more likely to change initial answer if it included disfluency (9.8% vs. 2.1%)
  - Suggesting disfluencies reflect response difficulty

Visual

- *R*s averted gaze (at least once) during more CIs (24.7%) than SIs (11.4%)
  - perhaps because they know conversational *Iwer*s can provide help, based on cues
- *R*s more likely to change initial answer if averted gaze while answering (24.7% vs. 4.3%)
  - Suggesting gaze aversion reflects comprehension difficulty (or response difficulty more generally)

Implications:

- Gaze aversion works very much like disfluencies; *R*s display both when *Iwers* can help

Outline

- Interviews
  - Response time can be affected by characteristics of items, respondents and interviewers
  - *R*’s speech can predict answers, comprehension difficulty, and response accuracy
  - *R*’s visual behavior (gaze aversion) can predict reliability of responses
- Web surveys
  - *R*’s inactivity can indicate difficulty and trigger help
  - *R*’s mouse movement can indicate comprehension difficulty
  - *R*’s multitasking (leaving survey and returning) can be related to item nonresponse
  - *R’s* very fast answers can be discouraged with interactive prompting

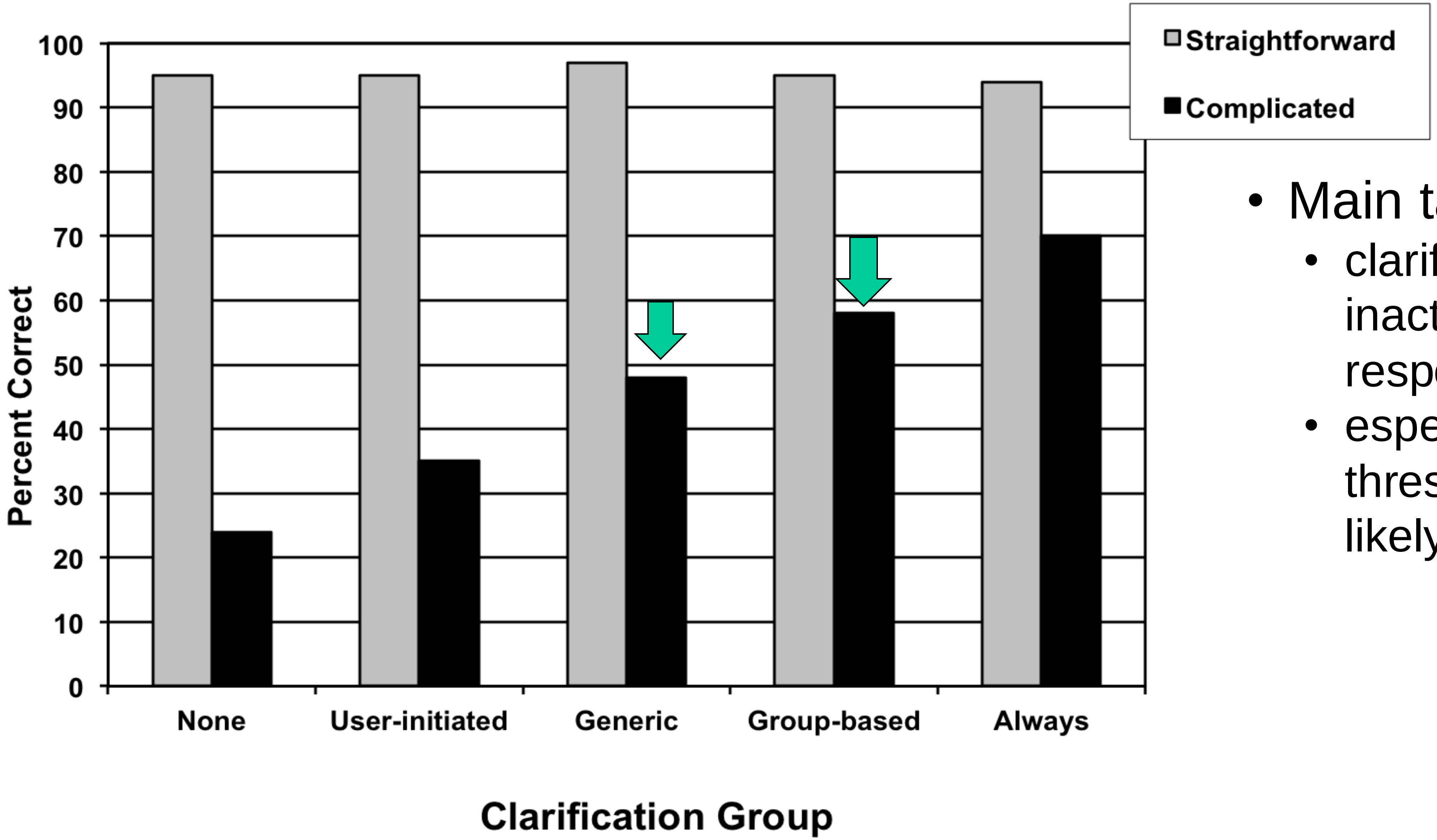
R’s Inactivity Can Indicate Difficulty and Trigger Help

- Online *R*s exhibit comparatively few types of paradata
  - primarily mouse movement and keypad entry – or their absence, i.e., inactivity
- Online requests for help likely to be rare
  - *R*s ask *Iwers* for help less often than need it (e.g., Schober & Conrad, 1997)
  - Web *R*s unlikely to use help features if require more than minimal effort (Conrad et al. 2006)
- Automatically presenting help if *R* seems to need it can reduce required effort/initiative to obtain it
- *Inactivity* is plausible indication of need for help (confusion/difficulty)
- Does automatically providing clarification when *R* is inactive improve response accuracy?

R’s Inactivity Can Indicate Difficulty and Trigger Help (2)

- Conrad et al. (2007) lab study (n=114) varied how clarification available:
  - None
  - R-initiated
    - *R must click hyperlinked terms*
  - Mixed initiative, generic
    - *R can click or the system can initiate clarification when R is inactive; same threshold for all Rs*
  - Mixed initiative, group-based
    - *Same as generic except longer inactivity threshold for older (72 yrs) than younger (26 yrs) Rs*
  - Always
    - *Definitions presented with all Qs for all Rs*
- Inactivity threshold:
  - 40% RT for straightforward questions in “None” condition;
  - in Mixed initiative, group-based condition, threshold set separately for young and old Rs

Response Accuracy by Clarification Group

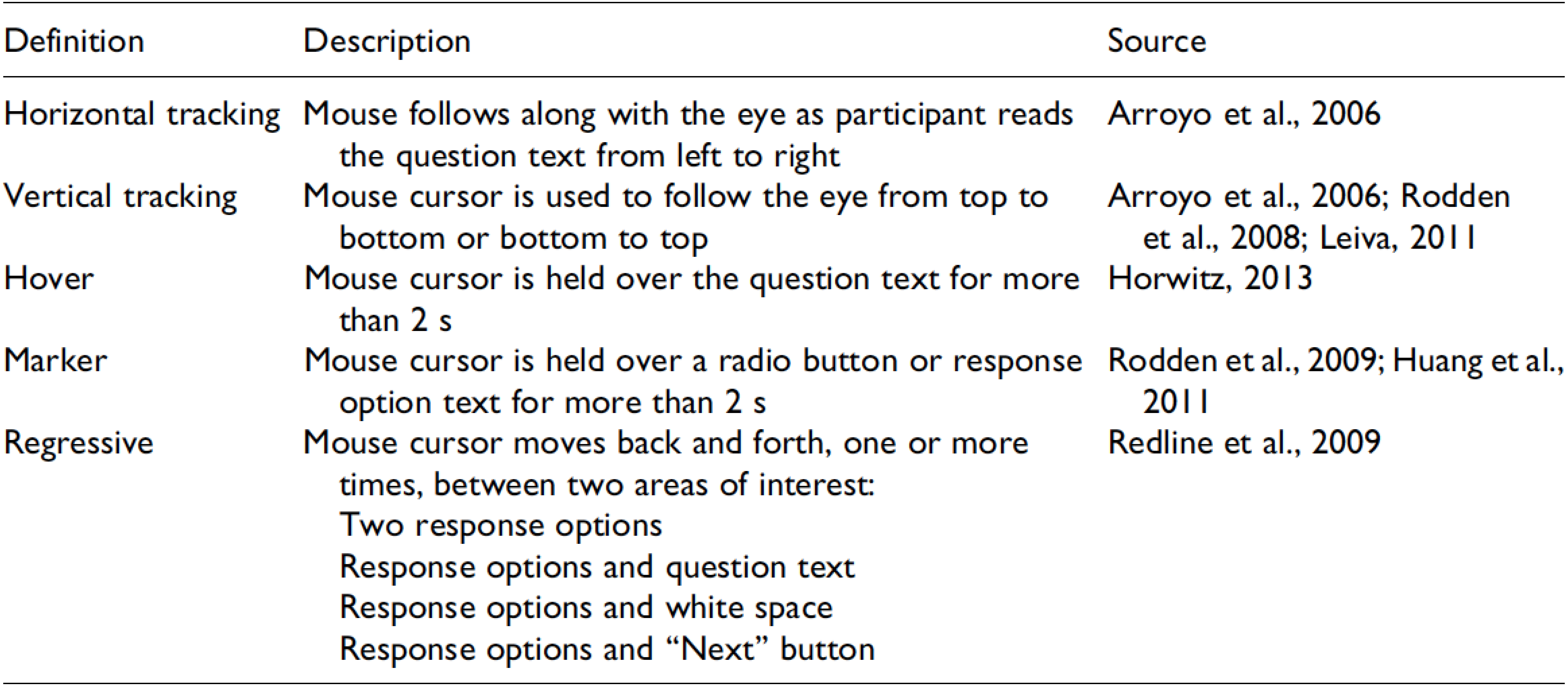


- Main takeaways:
  - clarification triggered by inactivity helps improve response accuracy
  - especially when inactivity threshold is tailored to *R*s' likely mental speed

*R'*s Mouse Movement Can Predict Response Difficulty

- Although *R* may be "inactive" with respect to clicking or typing, may move mouse in informative ways
  - Not generally visible in server-side paradata
- Horowitz et al. (2017) captured mouse movements in lab study of 100 *R*s
  - answered *Q*s based on straightforward and complicated scenarios

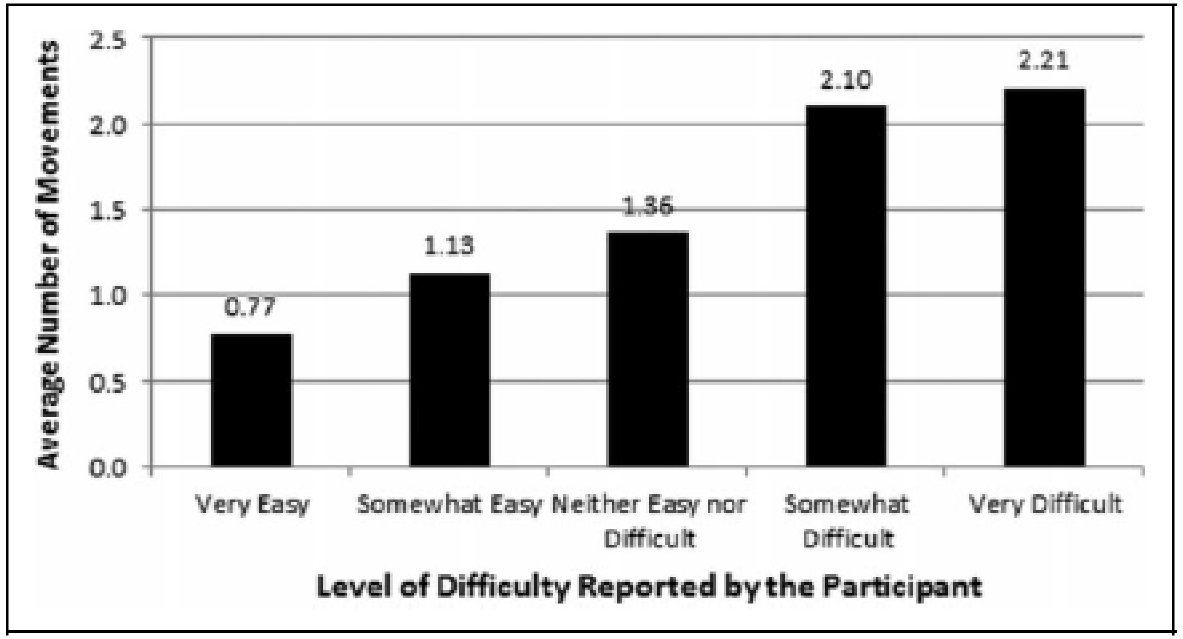
Types of Mouse Movement



Definition	Description	Source
Horizontal tracking	Mouse follows along with the eye as participant reads the question text from left to right	Arroyo et al., 2006
Vertical tracking	Mouse cursor is used to follow the eye from top to bottom or bottom to top	Arroyo et al., 2006; Rodden et al., 2008; Leiva, 2011
Hover	Mouse cursor is held over the question text for more than 2 s	Horwitz, 2013
Marker	Mouse cursor is held over a radio button or response option text for more than 2 s	Rodden et al., 2009; Huang et al., 2011
Regressive	Mouse cursor moves back and forth, one or more times, between two areas of interest: Two response options Response options and question text Response options and white space Response options and "Next" button	Redline et al., 2009

Mouse movements predict difficulty

# movements increases with *R*-rated *Q*-difficulty



Most predictive movements

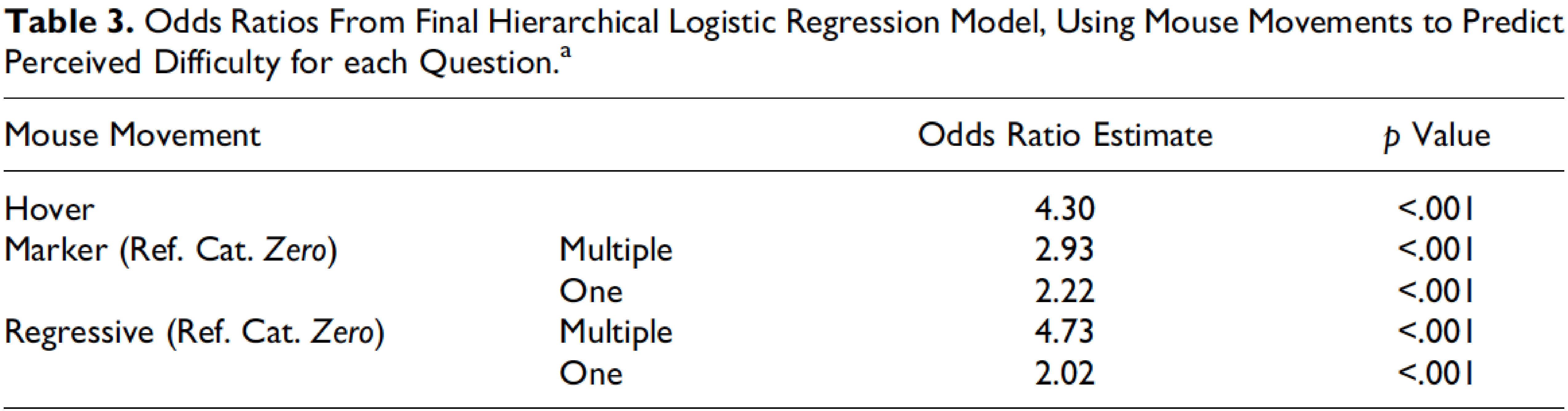


Table 3. Odds Ratios From Final Hierarchical Logistic Regression Model, Using Mouse Movements to Predict Perceived Difficulty for each Question.[a]

Mouse Movement		Odds Ratio Estimate	*p* Value
Hover		4.30	<.001
Marker (Ref. Cat. *Zero*)	Multiple	2.93	<.001
	One	2.22	<.001
Regressive (Ref. Cat. *Zero*)	Multiple	4.73	<.001
	One	2.02	<.001

Multitasking by Web *R*s

- Multitasking is widespread in browser use, especially switching between tabs
- Thus, multitasking by web survey *R*s may be widespread
- To what extent does respondent multitasking (RM) harm response quality?
- Sendelbah et al. (2014) examined prevalence of RM using two kinds of paradata:
  - *focus-out* events: *R* leaves browser and returns
  - *time-out* events: *R* inactive for longer than a page-specific threshold
- and relation to two measures of response quality (RQ):
  - item nonresponse
  - non-differentiation

Multitasking by Web *R*s (2)

- Web survey of university students about satisfaction with exchange program; completed on computers, i.e., not mobile devices
  - n=267; median completion time = 10 mins
- RM reasonably prevalent:
  - 42% *R*s exhibited at least one focus-out event
  - 53% exhibited at least one time-out event
  - 62% exhibited at least one of either type
- Focus-out count significantly (*p* = .04) predicts item nonresponse; time-out count marginally (*p* = .08) predicts item nonresponse
- No relationship between either RM indicator and non-differentiation
- Authors conclude that despite RM prevalence not major threat to RQ

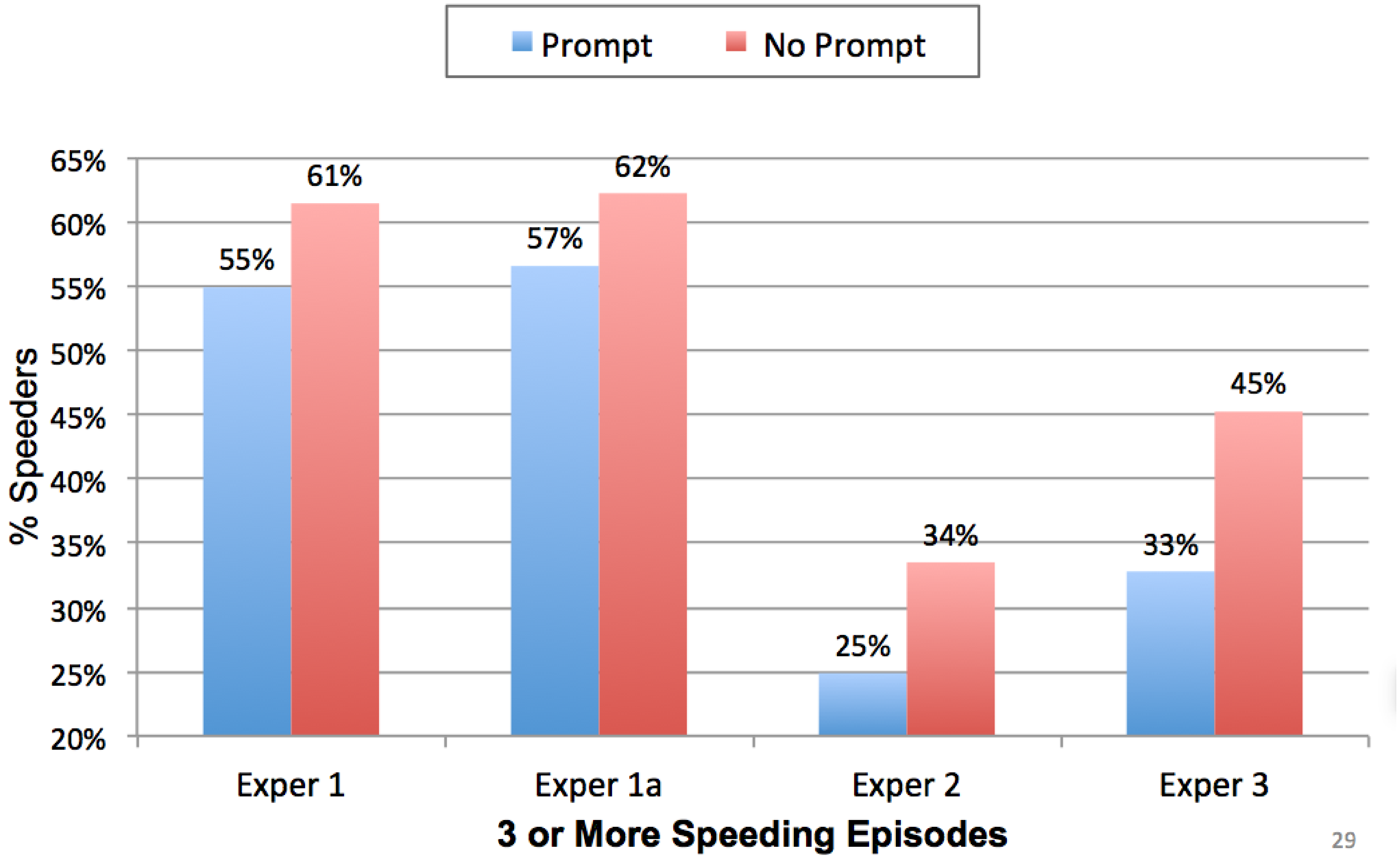
Using RT to Trigger Intervention: Speeding

- Prompting *R*s caught speeding (answers < 300 msec per word) in web surveys slows subsequent responses (Conrad et al. 2017)

  > *You seem to have responded very quickly. Please be sure you have given the question sufficient thought to provide an accurate answer. Do you want to go back and reconsider your answer?*

- 7 *Q*s about autobiographical quantities (for which true value not known)
- 7 simple arithmetic/probability questions for which true value known (in one experiment); enables measurement of response accuracy

Effect of Interactive Interventions on Speeding



- Prompting speeders to slow down was effective
- Improved response accuracy in arithmetic/probability problems for some Rs (in particular, Rs with some college or assoc. degree)

Paradata in Measurement: Summary/Conclusion

- Byproducts of response process, i.e., respondent paradata, anticipate response quality
  - Interviews: RT, Spoken paradata, Visual paradata
  - Web surveys: Inactivity, Mouse movements; Multitasking; Speeding
- In principle all of these can be used to trigger interventions and help overcome response difficulties and undesirable *R*-behavior
  - Intervention has been explored primarily in online data collection with promising results
  - In interviews, interview software can potentially track *R*s' disfluencies, pauses/RTs and prompt interviewer to offer help
- Possible next steps:
  - Develop tools to alert interviewers to verbal paradata indicating response difficulty
  - Test intervention approach in production surveys, both online and *Iwer*-administered