2023 · AMAZON ALEXA

Voice First Multimodal Design

Do’s and Don’ts

LISA FALKSON, SENIOR CONVERSATION DESIGNER
SEPTEMBER 2023

Agenda

“Everyone has a hidden agenda.
Except me!”

-Michael Crichton

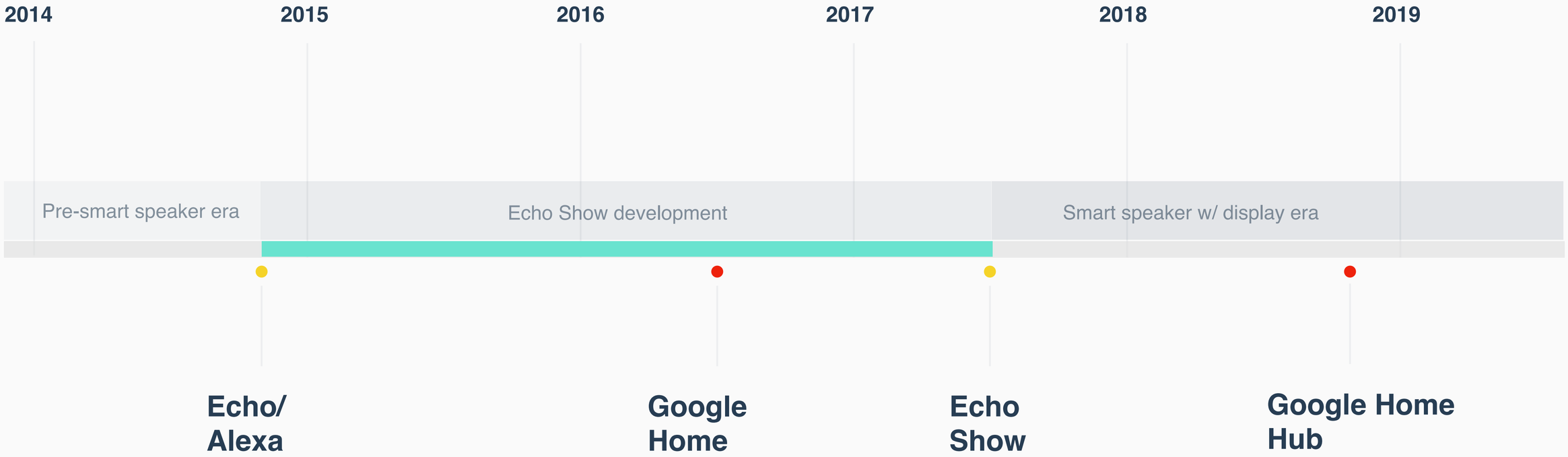
1. **Brief History: from “headless” to multimodal**
2. **Cognitive Load**
3. **Visual Dominance**
4. **Timing (is Everything)**
5. **Real World Examples**
6. **Conclusion**

Brief History

“I believe that the more you know about the past, the better you are prepared for the future.”

-Theodore Roosevelt

Timeline


2014
2015
2016
2017
2018
2019
Pre-smart speaker era
Echo Show development
Smart speaker w/ display era
Echo/ Alexa
Google Home
Echo Show
Google Home Hub

“There’s no reason to have an Echo with a screen except for video calls (with Alexa Comms).”

— A FRIEND OF MINE

“The Echo Show does everything the original Echo does, but many of those features are much better when you add in a screen.”

— ENGADGET, JUNE 2017

Examples

Visual output can improve the experience for many use cases

- **Question & Answer**

  “Alexa, what’s the weather?” (details)
  “Alexa, what does an elk look like?”

- **“Show me”**

  “Alexa, show me pictures of brownies”

Cognitive Load

“You talk too much
You never shut up
I said you talk too much…”

-Run-D.M.C.

Speech & audio are ephemeral

Smart assistant responses and follow-up questions are generally short and to-the-point, because:

- Working memory can only hold around **3-4 bits of information** at one time
- People tend to lose focus and attention after **12-15 seconds**
- Simple vocabulary and short strings minimize overall cognitive load
- Text-to-speech (TTS) is less comprehensible than human speech, produces “ear fatigue”

Speech + visuals can be better, or worse!

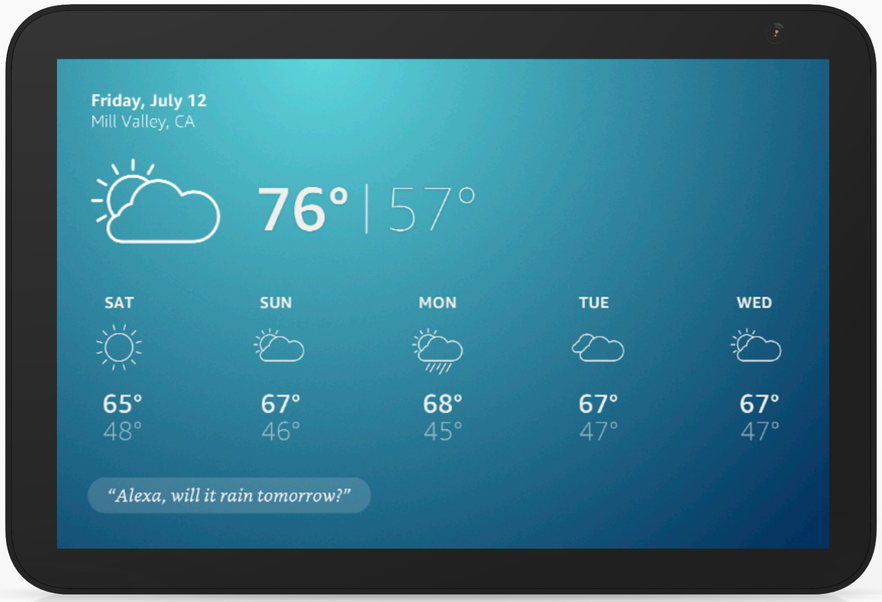
Better when….

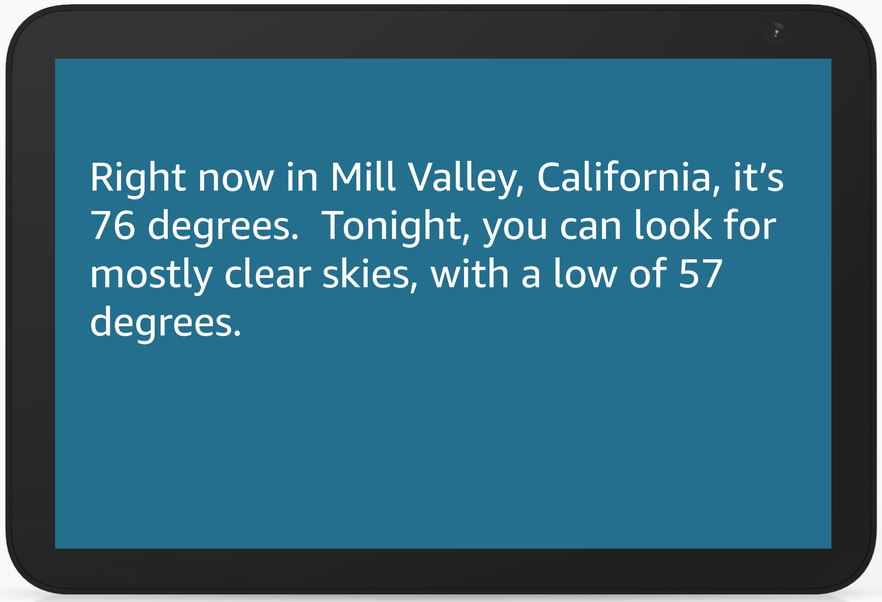
- Visuals enhance the experience: photos and illustrations
- On-screen text is short
- Text highlights the key action described in audio, placed next to graphic

Worse when…

- Screen shows large blocks of text
- Text on screen is identical to audio

“Alexa, what’s the weather?”

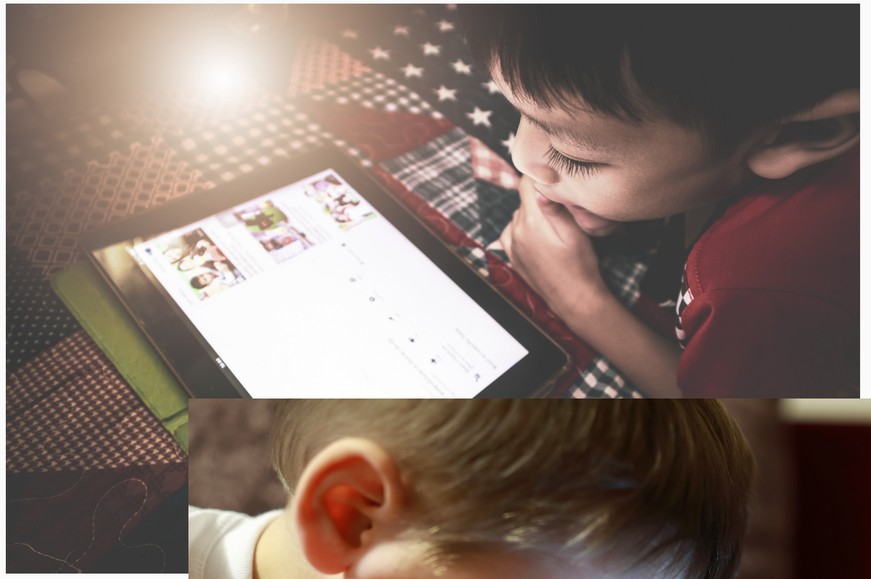


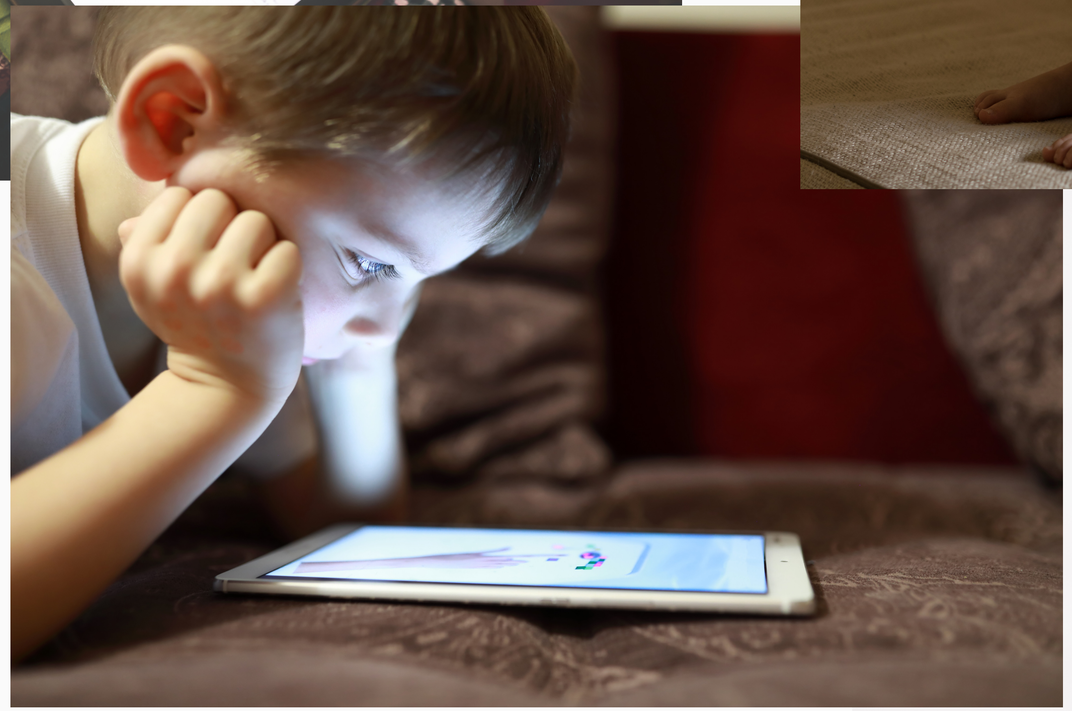


Visual Dominance

“Better to see something once than to hear about it a thousand times.”

-Asian Proverb





Visual + Auditory

When audio is accompanied by visuals, the visuals “dominate”

- People can’t hear/pay attention to audio when full text is on-screen
- Consider “the redundancy effect” in education

—> Students learn better from multimedia lessons containing graphics and narration than from graphics, narration, and redundant on-screen text[1]

[1]Mayer, R.E., Moreno, R. (1974).

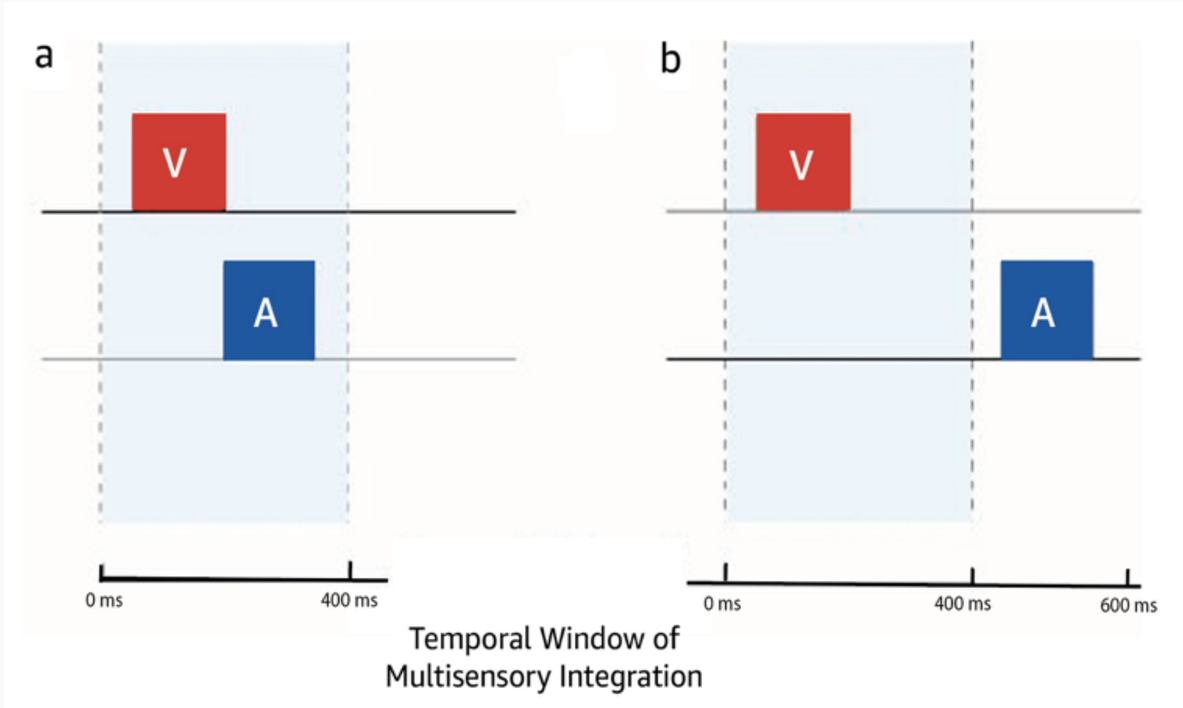
Timing (is Everything)

“It’s always about timing. If it’s too soon, no one understands. If it’s too late, everyone’s forgotten.”

-Anna Wintour

Temporal Binding Window

Visual and auditory events within approximately 400 ms are considered simultaneous



Temporal Window of
Multisensory Integration

Temporal Binding Window

Outside this window, audio and visuals are processed separate events

Order and timing are important!

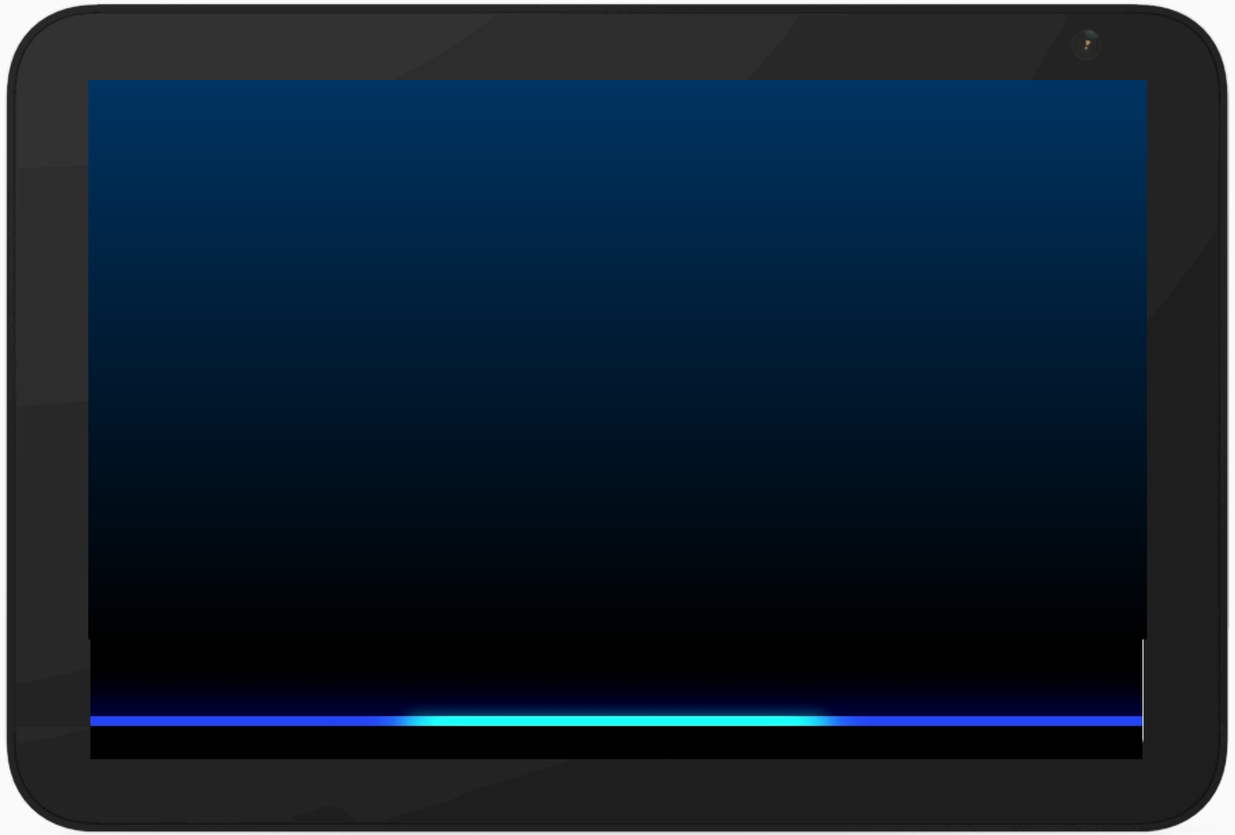
“Alexa, turn on the TV”

Real World Examples

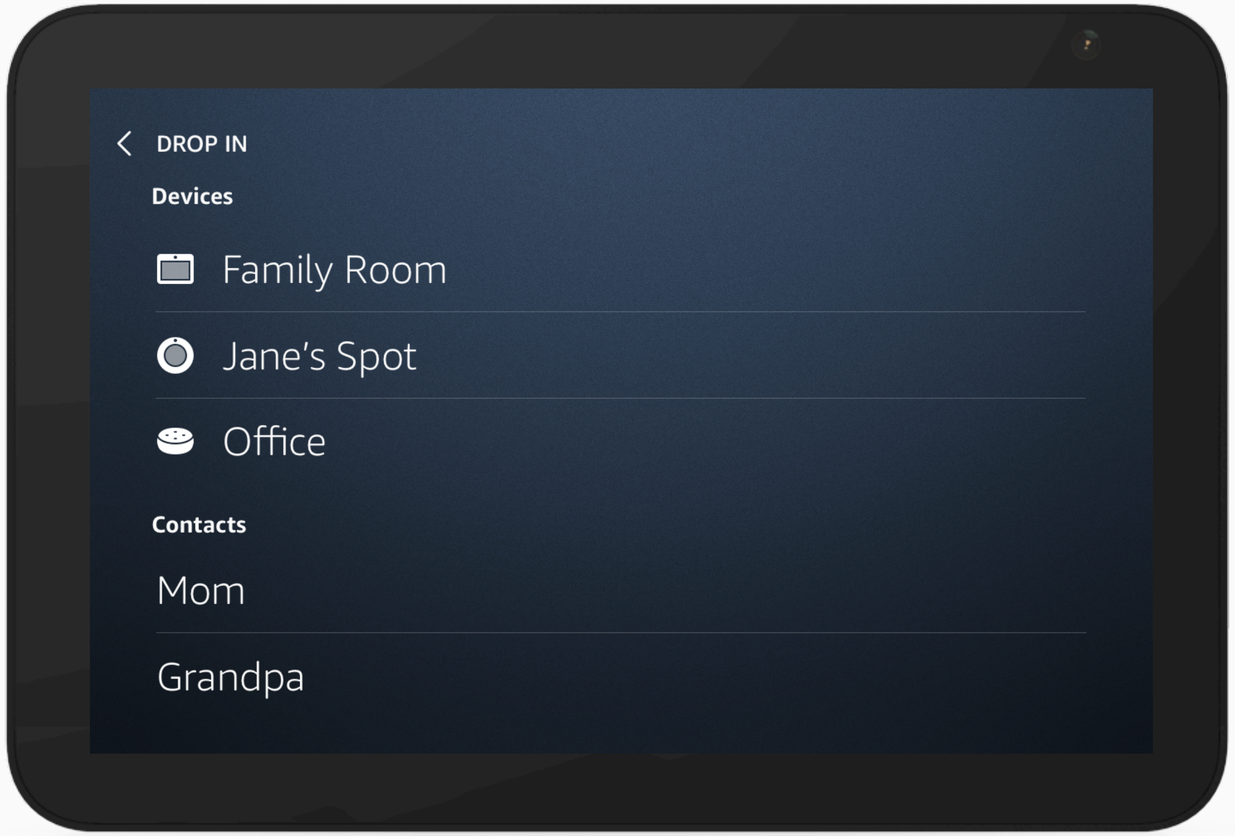
“Nothing happens in the ‘real’ world unless it first happens in the images in our heads”

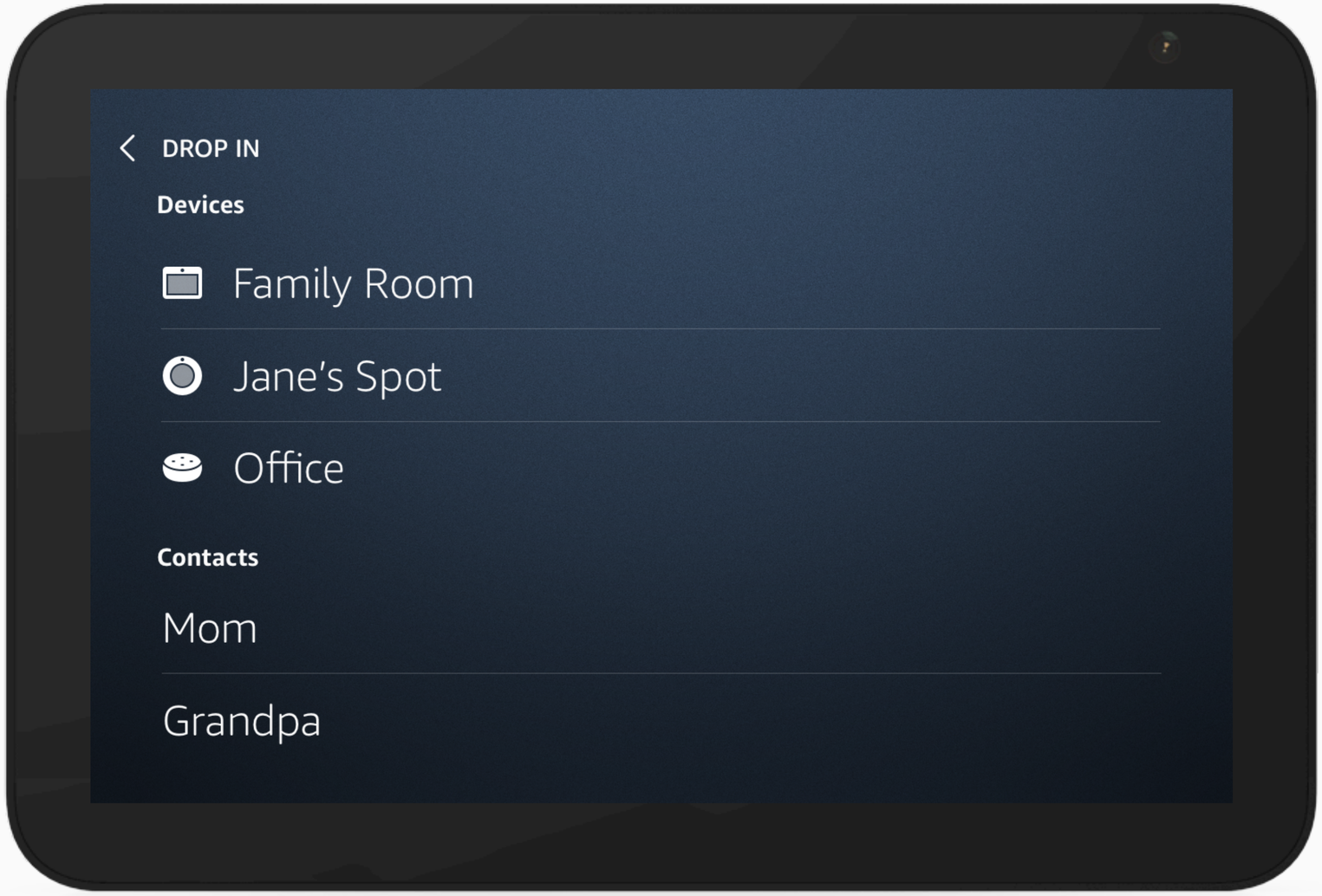
-Gloria E. Anzaldúa

“Alexa, Drop In”



“Alexa, Drop In”





01

Drop In MMD

Positive results when this screen was added to the audio-only experience

- On screened devices, dialogs that feature multimodal disambiguations succeed 4.37% more often than VUI-only dialogs.
- Successful multimodal dialogs use touch ~18% of the time.

Designs for Calling

Who's calling: Robbie or Alyse?

What's the phone number?

Katherine or Catherine?

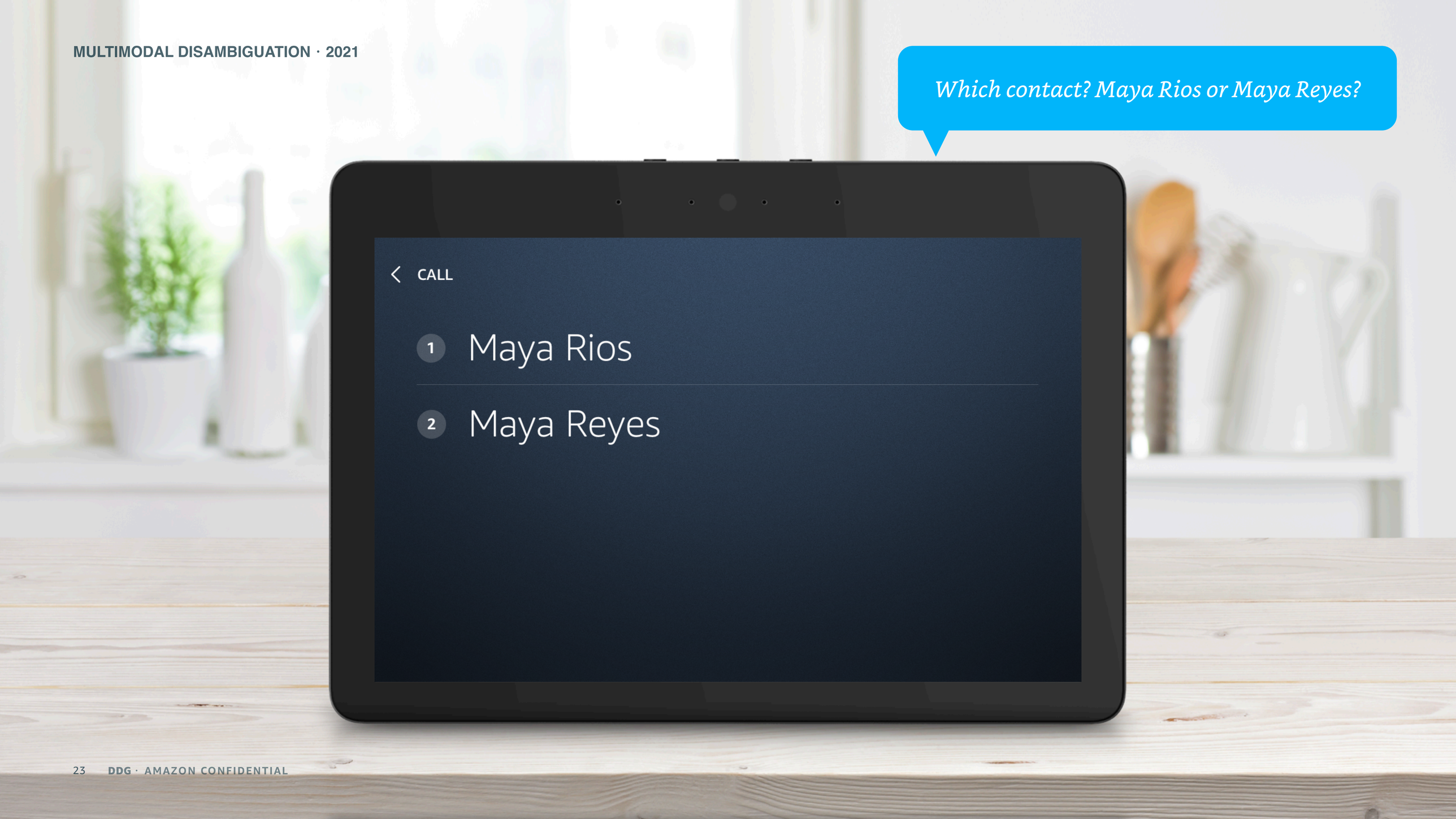
I can reach Steve Adams' Alexa devices or phone. Where should I call?

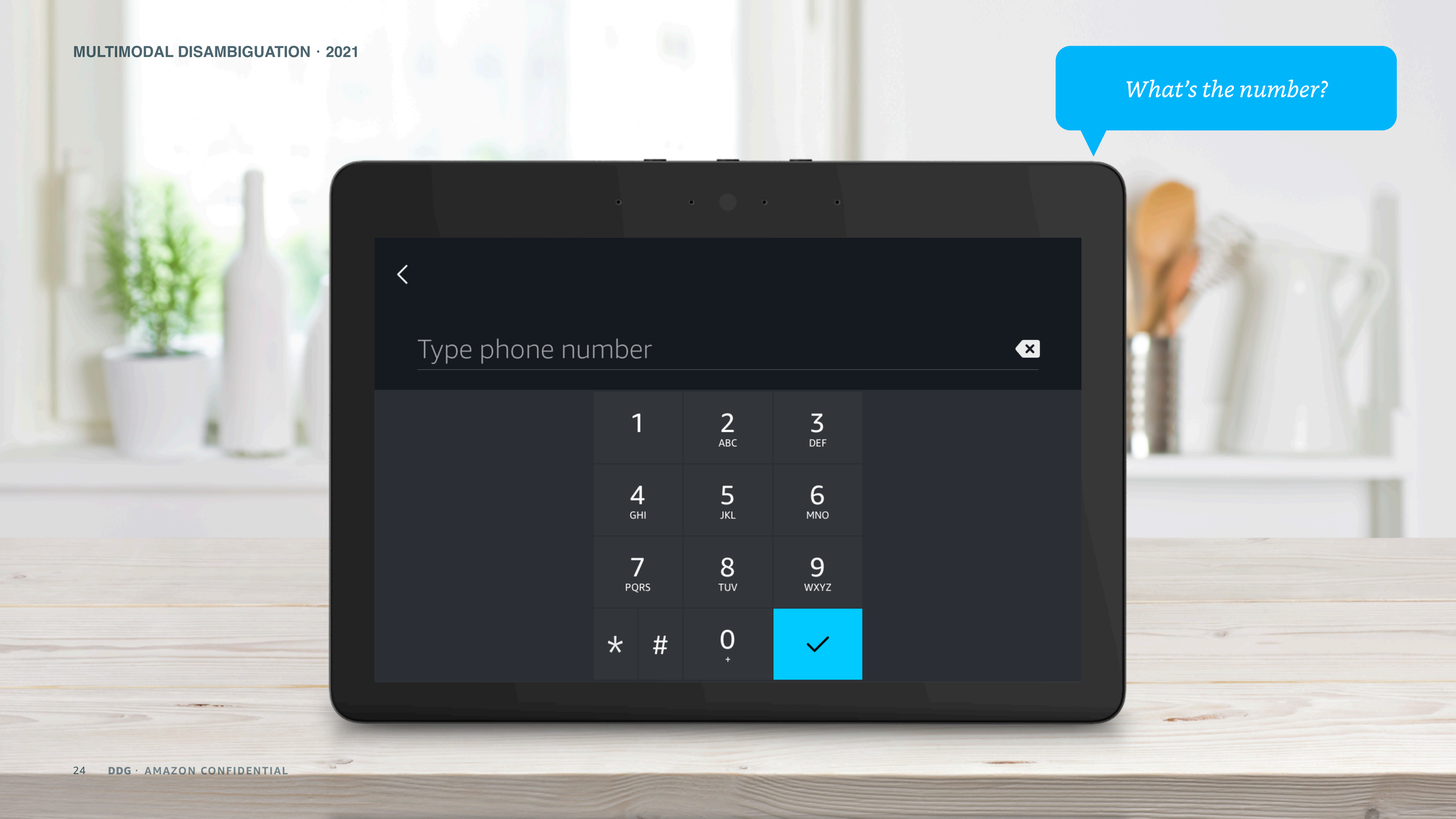
Which phone number, contact or device do you want to call?

I heard you ask for Mark Cuban but I can't find Mark in your contacts do you know the phone number?

Call Andrew's work or home phone?

Steve Adams, right?


Which contact? Maya Rios or Maya Reyes?
CALL
1 Maya Rios
2 Maya Reyes


What's the number?
Type phone number
1
2 ABC
3 DEF
4 GHI
5 JKL
6 MNO
7 PQRS
8 TUV
9 WXYZ
* #
0 +

Relevant Findings from Testing

- Most participants continued to use voice. When questioned most understood they could touch, but continued in the voice because they considered Echo to be a voice first device.
- Too much displayed information increased cognitive load and seemed to overwhelm participants.
- Adding buttons to the screen did NOT influence whether they continued with voice or touch, but the text we put on the screen DID influence what they chose to say in response to the prompt.

Conclusion

“I liked it, it was good, but I feel like what was missing was a **call to action** at the end.”

-Random guy at one of my talks in 2018

Do

- Design **explicitly** for multimodal instead of simply adding visuals like text, photos
- Run additional usability studies for multimodal devices. Things we've found:

  —> Screen adds new headings & numbers? People will say them!
  —> Consider visual navigation commands ("next," "go back", "home")
  —> List items: visual reference or tappable buttons?

- Look where voice-only **fails**, consider how screen support could help:

  —> Long list items
  —> Error recovery

Don’t

- Overwhelm the customer with large blocks of text or overly complex visuals
- Assume that the voice-only and multimodal experiences will be the same.
- Mirror exact TTS text as on-screen text

Call to Action

- Evaluate your existing voice-only vs. voice + screen interactions, end-to-end
- See where you can optimize:

    1) **Minimize cognitive load** —> visuals compliment audio
    2) **Consider visual dominance** —> avoid masses of text in “karaoke mode”
    3) **Review timing of audio & visual cues** —> adjust for temporal binding window

Thanks!

Lisa Falkson

https://www.linkedin.com/in/lisafalkson/
@LisaFalkson

References

- Colavita, F. B. (1974). Human sensory dominance. Perception & Psychophysics, 16(2), 409–412. https://doi.org/10.3758/BF03203962
- Mayer, R.E., Moreno, R. (1974). Nine Ways to Reduce Cognitive Load in Multimedia Learning, Educational Psychologist http://faculty.washington.edu/farkas/WDFR/MayerMoreno9WaysToReduceCognitiveLoad.pdf