CONSTRUCTION & EVALUATION OF CONTEXTUALLY-TAILORED ACTIVITY SUGGESTIONS: EVIDENCE FROM THE HEARTSTEPS PILOT

Kelly L. Hall¹,²; Timothy NeCamp³; Shawna N. Smith, PhD²,³; Gaurav Paruthi³; Mark W. Newman, PhD⁴; Susan A. Murphy, PhD¹,²; Predrag Klajšna, PhD³,⁴,⁵
¹University of Michigan Department of Statistics, ²UM Institute for Social Research, ³UM Division of General Medicine, ⁴UM School of Information, ⁵Group Health Research Institute

Objective
Use crowdsourcing tools to create a library of contextually-tailored activity suggestions for HeartSteps and assess participant satisfaction and message effectiveness following a six-week pilot study.

Background
Tailored health communication is effective at encouraging health behavior change1⁴, but most tailoring to-date has focused on static and personal, as opposed to dynamic and contextual, characteristics.

In just-in-time adaptive interventions (JITAs)⁵, which adapt intervention delivery to context, contextual tailoring may lower barriers to treatment buy-in and adherence.

Passive sensing tools makes such tailoring feasible, but also change the scale of tailoring required. JITAs require libraries of messages tailored for all combinations of important contextual variables.

Expert-guided crowdsourcing pipelines provide one solution for scaling tailored message creation.

Crowdsourced Message Construction
Amazon’s MTurk was used to crowdsource message creation. Workers were given one of two tasks:

1. Generate messages for a specific persona in a specific context, following some simple writing suggestions.
2. Rate messages created by other users, from 0 to 9.

295 messages were generated and rated by 5-10 Turkers before delivery to research team.

Fig. 2 Rating distribution from Mechanical Turk users.

Expert-Guided Curation
The goal of curation was to create a library of messages representing health communication ‘best practice’. 30 messages created by a UM health communication expert served as our guide.

MTurk messages rated 6 or higher were edited to ensure:

1. Messages were in line with health communication best practice (e.g., encourage, not dictate; suggestions, not commands)⁶-⁹
2. Specific and actionable phrasing.
3. No exclusive language (e.g., references to specific exercise equipment, spouse, children)
4. A cohesive “voice” for the application.
5. 200 characters or fewer, no spelling/grammar errors.

New messages were also written to ensure coverage of all combinations of context. In total, over 500 messages were generated.

Messages were “tagged” for context (location, time of day, day of week, and weather), resulting in 90 context “buckets.” Messages could fit in more than one bucket. Each bucket had at least 30 messages.

Fig. 1 Examples of HeartSteps activity suggestions.

Results
Messages were used in a 6-week pilot study with 37 participants. Participants rated suggestions with thumbs up or down upon receipt.

Descriptive Statistics
• 2749 suggestions were delivered to participants for which responses were recorded.
• 1350 (49.1%) were rated thumbs up, 529 (19.2%) thumbs down, and 870 (31.6%) no response.

Mixed Effects Model
• Used random-effects model with an autoregressive 1 correlation structure to model logged step count in the 30 minutes following message delivery by response (thumbs up, down, or no response)
• Messages rated thumbs up were associated with step count increases by a factor of 1.77 (~163 steps in next 30 min) compared to thumbs down and 2.62 (~343 steps in next 30 min) compared to no response
• Both of these differences significant at p < 0.01

Qualitative Feedback from Exit Interviews
1. More variety
   • Need for more dimensions (e.g. activity type) and deeper tailoring
2. Heterogeneity in user communication preferences:
   • Some wanted a buddy, others wanted a coach
3. Double-tension of tailoring:
   • “Intimate” relationship formed; when context/suggestion is wrong, this trust is breached and sometimes difficult to recover

Next Steps
Future versions of HeartSteps will be up to 12 months long, necessitating a large library of actionable, contextually-appropriate messages:

1. How can we improve the pipeline to minimize expert editing time?
2. What instructions and scenarios yield the best MTurk results?
3. How extensive does our message library need to be, and how often will it need to be refreshed?

Contact
Kelly L. Hall
Research Assistant,
Institute for Social Research
University of Michigan
kellyhal@umich.edu
https://sites.google.com/a/umich.edu/kellyhall/

References