Check for updates

Empirical Research

Journal of Mixed Methods Research
2022, Vol. 0(0) 1–23
© The Author(s) 2022

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/15586898221096934
journals.sagepub.com/home/mmr


Persuasion with Precision: Using Natural Language Processing to Improve Instrument Fidelity for Risk Communication Experimental Treatments

Ann Marie Reinhold[1,3,4], Eric D. Raile[2], Clemente Izurieta[3], Jamie McEvoy[4,5], Henry W. King[3], Geoffrey C. Poole[1,4], Richard C. Ready[4,6], Nicolas T. Bergmann[5], and Elizabeth A. Shanahan[2]

Abstract

Instrument fidelity in message testing research hinges upon how precisely messages operationalize treatment conditions. However, numerous message testing studies have unmitigated threats to validity and reliability because no established procedures exist to guide *construction* of message treatments. Their construction typically occurs in a black box, resulting in suspect inferential conclusions about treatment effects. Because a mixed methods approach is needed to enhance instrument fidelity in message testing research, this article contributes to the field of mixed methods research by presenting an integrated multistage procedure for constructing precise message treatments using an exploratory sequential mixed methods design. This work harnesses the power of integration through crossover analysis to improve instrument fidelity in message testing research through the use of natural language processing (NLP).

Keywords

message construction, Narrative Policy Framework, validity and reliability, exploratory sequential mixed methods, crossover analysis

[1]Department of Land Resources & Environmental Sciences, College of Agriculture, Montana State University, Bozeman, MT, USA
[2]Department of Political Science, College of Letters & Science, Montana State University, Bozeman, MT, USA
[3]Department of Computer Science, Gianforte School of Computing, Montana State University, Bozeman, MT, USA
[4]Montana Institute on Ecosystems, Montana State University, Bozeman, MT USA
[5]Department of Earth Sciences, College of Letters & Science, Montana State University, Bozeman, MT, USA
[6]Department of Agricultural Economics & Economics, College of Agriculture, Montana State University, Bozeman, MT, USA

Corresponding Author:
Ann Marie Reinhold, Department of Land Resources & Environmental Sciences, College of Agriculture, Montana State University, P.O. Box 173120, Bozeman, MT 59717-3120, USA.
Email: annmarie.reinhold@montana.edu

Message testing research investigates the persuasive power of different kinds of messages. Across multiple fields (e.g., health science, marketing, public policy, political science, hazard preparedness, and environmental risk communication), much qualitative, quantitative, and mixed methods research focuses on understanding how messages influence attitudes, beliefs, and behaviors (Ajzen, 1991; Fishbein, 2008; National Cancer Institute, 2005). While the core component of many of these studies is testing the persuasive power of messages, there is surprisingly little transparency and guidance on how researchers should develop message treatment conditions, which are randomly assigned to participants to determine the effects of treatments of different messages against a control condition. As noted by Onwuegbuzie et al. (2010, p. 57), "scant guidance has been given to help researchers use mixed research techniques to optimize the development of either qualitative or quantitative instruments." A methodological gap exists in message testing research because a procedure guiding the *construction* of message treatments to optimize their instrument fidelity is lacking (see Table 1 for definitions of terms used throughout the manuscript). While a handful of mixed methods scholars (Howell Smith et al., 2020; Khanal, 2013; Onwuegbuzie et al., 2010) detail procedures of how mixed methods can optimize instrument fidelity through data integration regarding survey questions, we expand on this body of work by specifying how mixed methods can improve instrument fidelity of treatment conditions in experimental testing.

In experimental message testing research, precise messages operationalize intended treatment conditions. A poorly constructed message can result in Type I errors—rejecting a true null hypothesis—if a message tests unintentional or tangential concepts treatments and Type II errors—accepting a false null hypothesis—if the constructed message dilutes the treatment condition. Thus, the instrument fidelity imparted by precise message construction is the foundation for sound investigation. Common strategies used in message construction include needs assessment focus groups, participant phrasing discussions, and interviews with target audiences (Willoughby & Brickman, 2020). However, the instrument fidelity of messages is suspect because no well-established, transparent procedures for scientific message construction exist despite the fact that scholars in these fields expend tremendous effort in describing and analyzing participant responses to messages. While there is an abundance of important theoretical discussions of mixed

Table 1. Definition of Terms.

Component	Definition	Resource
Precision	Exactness of treatments as demonstrated through theoretical grounding, validity, reliability, and instrument fidelity	Authors' definition
Instrument fidelity	"[M]aximizing the appropriateness and/or utility of the instruments used, whether quantitative or qualitative"	Onwuegbuzie et al., (2010, p. 57)
Theoretical foundation	Measurement and testing flow deductively from categories and characteristics described in established theoretical frameworks	Authors' definition
Ecological validity	"Whether or not the experimental situation captures the critical aspects of the real-world environment assumed to be important"	Schmuckler (2001, p. 430)
Construct validity	Extent to which an instrument or test reflects the particular construct being investigated	Cronbach & Meehl (1955)
Internal validity	Extent to which study conclusions about cause-and-effect relationships are sound	Bryman (2016, p. 41, p. 41)
Reliability	"Consistency of measurement over time or stability of measurement over a variety of conditions"	Drost (2011, p. 108)

methods standards of validity and data integration (Dellinger & Leech, 2007; Fàbregues & Molina-Azorín, 2017; Fetters et al., 2013), we address a methodological gap in these works regarding such standards in experimental research. Indeed, the current approaches to constructing messages occur in a black box, reliant largely on expert opinion, with the often unstated assumption that message construction accurately represents treatment conditions. With no established procedures to maximize instrument fidelity, many existing studies—including some of our own—engender real threats to validity and reliability.

The vagueness of message construction typically resides in how researchers incorporated data from the formative research phase into the messages that were subsequently tested. The following three studies are examples of how messages are constructed in message testing research that lack instrument fidelity. In the first example (Poehlman et al., 2019), messages were developed and tested that would "resonate with and inspire priority groups to act" on Zika virus prevention in Puerto Rico (p. 900). The research team first conducted qualitative, formative research with women in the Women, Infants, and Children program. They then conducted "environmental scans" to quickly collect information from a variety of publicly available sources. Finally, the team brainstormed concepts for their messaging campaign. Yet, *how* the researchers integrated the data from formative research and environmental scans into the messages is not described. Another example is a study (Hennink-Kaminski & Dougall, 2009) on messaging about the normalcy of infant crying, whereby researchers used findings from focus groups to develop "five broad creative approaches" that were presented to the leadership team. From these, two campaign concepts emerged. One of these concepts (i.e., "advice") included a "photo and quote from a real parent in North Carolina" (p. 58). However, no specific methodological details are provided for *how* the data from the focus group was integrated into the campaign concepts. Additionally, no method or procedure was given for how researchers selected a particular quote for use. In the third example (Barbour et al., 2015), the researchers discuss the importance of the design and structure of their messages, but using the passive voice, simply assert that "[m]ultiple messages were created to represent the experimental manipulations" (p. 818). Such a black box approach to message construction can threaten the validity of a study.

In an effort to address validity, some empirical studies rely on a textual basis for constructing messages. For example, one study adapted publicly disseminated political emails (McLaughlin et al., 2019). Another excerpted from actual news (McLaughlin, 2020), while another used segments from a television program (Semmler & Loof, 2019). Shanahan et al. (2014) used ideas from public comments, and others found passages from interviews (Leshner et al., 2018). Finally, several studies noted that questions asked during the focus group or other formative research phases were informed by theory (Jordan et al., 2012; Lapka et al., 2008). Despite these efforts, none of the studies specifically or explicitly described *how* the data from prior phases of the research project was integrated into the message construction phase for further testing.

While these studies have advanced our knowledge of the effects of messages (both science-based and narrative-based messages), our procedure takes a first step toward precision in message construction through the deployment of a systematic mixed method to develop message treatments (Figure 1). Specifically, our mixed methods procedure integrates data collection and data analysis (Fetters & Molina-Azorin, 2017) with an exploratory sequential mixed methods design (Fetters & Freshwater, 2015).

The aim of this article is to present our unique contribution to mixed methods research that addresses the methodological gap in message testing research: the need for a formalized procedure guiding the *construction* of message treatments to optimize their instrument fidelity. We work towards closing this gap through the development of a procedure that integrates qualitative semi-structured interview data and quantitative natural language processing (NLP) techniques for more

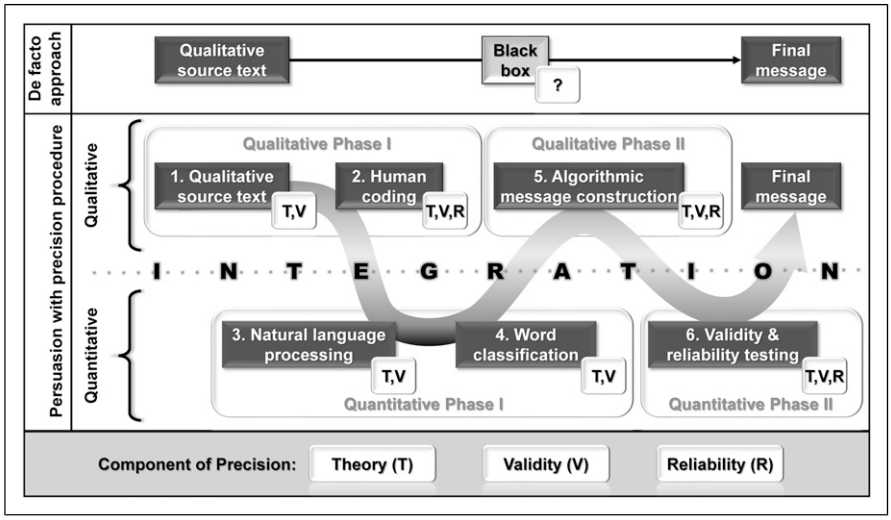

Figure 1. The de facto approach to message construction versus our integrated mixed methods procedure to improve the precision of experimental message treatment conditions. Upper panel: current "black box" approach to message construction that risks ineffective or erroneous operationalization of theory and relies heavily on expert opinion to mitigate threats to validity and reliability. Lower panel: Each qualitative and quantitative phase and step in our procedure is intended to improve a component of precision (Table 2). The wide arrow traversing the dotted line indicates points of integration; for example, the product of Qualitative Phase I was human-coded text, which was *integrated* into Quantitative Phase I as it became the basis of the natural language processing and word classification steps.

precise message construction for subsequent measurement and testing via focus groups, surveys, or other research methods.

The remainder of this article is organized as follows. The section *Shining a Light into the Black Box* illuminates the reasons for employing a mixed message approach to improve message construction and introduces the Persuasion with Precision Procedure (herein referred to as "the procedure"). The subsequent section introduces our *Exemplar*, the specific instance of a case study to which the procedure detailed here was applied. Because the exemplar focused on narrative messages, the following section, *Why Narrative Messages?*, describes the reasons for employing narratives when messages are constructed for persuasive purposes. The subsequent section, *Why NLP?*, introduces this suite of computational tools, highlights their utility in the digital humanities, and describes why we employed them in the procedure. Next, the *Detailed Procedure* section is a step-by-step description of the procedure and how it was applied to our specific exemplar. Finally, the *Discussion* focuses on how the procedure contributes to the field of mixed message research, concluding with a subsection devoted to limitations and future directions.

Shining a Light into the Black Box: Why Employ Mixed Methods to Improve Precision in Message Construction?

We asked the question: *how can we build messages that precisely operationalize treatment conditions?* Our answer to this question directly reflects our position that a mixed methods approach is the research frontier for improving precision in message construction for research purposes. Using an exemplar of a case featuring riverine flooding, we present an integrated

multistage procedure for constructing precise messages using an exploratory sequential mixed methods design (Fetters & Freshwater, 2015). Each step in this procedure shines light into the black box of message construction, with the specific purpose of improving precision (Figure 1; Table 2). By carefully describing each step in the procedure, we outline a path for surmounting the challenge of achieving integration (Bryman, 2007; Fetters et al., 2013; Uprichard & Dawney, 2016). Briefly, the qualitative steps ground the procedure in theory and impart validity and reliability in the final messages via employment of well-established techniques. In turn, the quantitative steps complement the qualitative steps in ways that systematically reduce threats to validity and reliability and improve the operationalizing of theory. Importantly, the integration between the qualitative and quantitative phases in this design improves precision of message treatment conditions.

Exemplar: Constructing Narrative Messages to Communicate Riverine Flood Hazard Information

In this article, we use narrative messages in a riverine flood hazard context as an exemplar to illustrate how mixed methods can achieve integration in a transparent, systematic manner. The empirical goal of our exemplar was to measure the power of narrative risk communication to influence the audience's affective response to a message (i.e., the valence and intensity of emotions) and their intended risk mitigation behaviors Raile et al. (2022) and Shanahan et al. (2019a). Specifically, we sought to learn how the narrative mechanism of "character selection" works to persuade in narrative science messages about flooding and whether narrative messages that highlight "hero" characters generate different affective responses and decisions than narrative messages that highlight "victim" characters. Therefore, building strong message treatments with different character sets was of paramount importance to our exemplar.

The Study Area of Our Exemplar

The Yellowstone River basin in Montana, USA, is keenly susceptible to flooding hazards. With a classic mountain-snowmelt hydrologic regime, the Yellowstone experiences frequent flooding events. These conditions are made especially acute when combined with higher frequencies of "rain on snow" events. Local hazard preparedness is vital to avoid the hazard-to-disaster trajectory. This iconic river originates in Yellowstone National Park and flows 1,100 KM northeast to its confluence with the Missouri River in western North Dakota. The Yellowstone is the longest unimpounded river in the conterminous 48 states and flows through several communities in Montana (Figure 2). Land on the Yellowstone River is held by private landowners and federal and state management agencies. The river is integral to many different communities for agricultural, residential, industrial, and recreational purposes. Without appropriate hazard preparedness, the increased volatility introduced by climate change will significantly increase the vulnerability of individuals and communities (Whitlock et al., 2017).

Why Narrative Messages?

At the heart of conventional risk communication is the assumption that scientific information on the probability and consequences of natural hazards will lead people to engage in risk-reduction behaviors (Ludy & Kondolf, 2012); however, many studies reveal that scientific information in isolation rarely affects hazard preparedness (Wachinger et al., 2013). New information about flood hazards is unlikely to prevent hazard-to-disaster trajectories because an alarming gap persists between scientific predictions of hazards and the general population's perceptions of risks

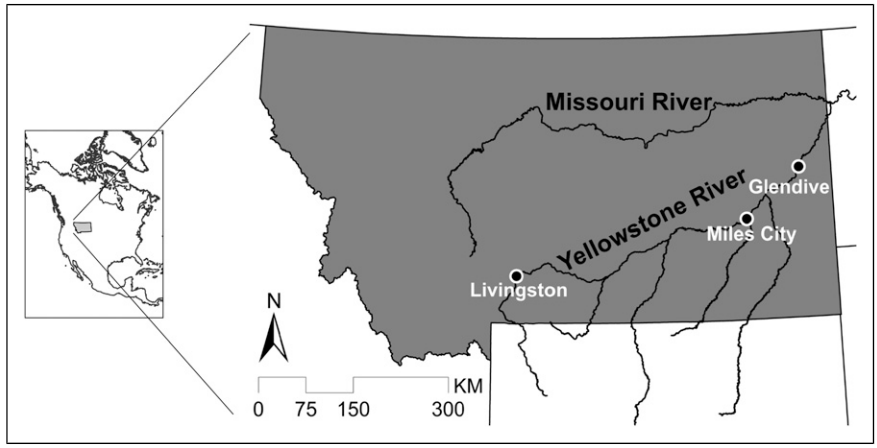

Figure 2. Location of the Yellowstone River in Montana, USA, and the three cities that compose the study area of the exemplar. Livingston, Miles City, and Glendive are all flood-prone cities in which semi-structured interviews were conducted to provide qualitative source text (Step 1 Qualitative source text) and in which message validity and reliability testing occurred (Step 6 Validity & reliability testing).

associated with those very same hazards (Barnes, 2002). In turn, preparedness decisions are often based on subjective factors derived from life experiences and cultural values rather than up-to-date science information (Bubeck et al., 2012). In our broader project, we propose that one way to improve risk communication is to use narrative structure to relay the story of the scientific information.

Why *narrative* messages? According to the Narrative Policy Framework (NPF) (Shanahan et al., 2018b), people communicate about and understand their world primarily through narratives or stories, filled with characters, plot, and action; as such, narratives are powerful in shaping opinions and decisions. Risk perceptions and hazard-mitigating decisions are continuously communicated and reified in narrative form between family members, neighbors, and communities. As described in the NPF, recognizable, stable, and observable structural elements (e.g., characters, moral of the story or decision) are inherent in narratives. The most vital element is that of the character, as it is foundational to what constitutes a narrative (Shanahan et al., 2013). Characters generally fall into three categories: victims, heroes, and villains. Thus, the type of characters cast in the narrative and their associated actions are formative in constructing different notions of reality and consequent decisions for the intended audience. Whereas narrative structures are stable, content varies across narratives—from flood to health hazards, for example. In our exemplar, we sought to understand the persuasive power of precise structural mechanisms in a risk communication context. However, the integrated multistage procedure we present here may be applied to other contexts such as public health and marketing.

Testing narrative-based risk communication is not new but has been consistently imprecise because black-box message construction impairs instrument fidelity. Inferring causal mechanism(s) from black-box messages is dubious; in particular, internal validity is compromised if messages lack ecological and construct validity, regardless of how well researchers measure and analyze the responses to those messages. While previous risk communication studies have examined the differences between the impact of technical information and narratively presented hazard information (Barbour et al., 2016; Occa & Suggs, 2016), the narrative treatments lacked validity, as they were ad hoc constructions made up by researchers. Our approach was to reduce

threats to validity by capturing narrative elements directly from residents in the study area to catalog in vivo local language used to describe flood hazards, a step referred to as participant enrichment (Collins et al., 2006).

Why NLP?

Natural language processing refers to a cadre of machine-learning tools that can be applied to infer, describe, and quantify the meaning and nuance in human language transcripts. NLP techniques enable efficient and unbiased processing of large bodies of texts—such as coded interview transcripts or other qualitative source texts—that would be unwieldy or impossible for a human to process manually. These techniques bring out qualities of narratives that are impossible to discern with the naked eye (Flanders, 2005). We assert that NLP strengthens the operationalization of the qualitative and theoretical foundations of our research procedure by enabling the identification and relative importance of the words that most precisely capture the treatment conditions from the source texts.

Application of computational science techniques in the subfields of digital humanities and computational social science are limited (Grubert & Siders, 2016). Yet, computational techniques have tremendous potential to help social scientists make valid causal inferences and develop theory from the assessment of large and unwieldy datasets (Grimmer, 2015). Nascent application of machine-learning tools such as text mining, sentiment analysis, word frequency analysis, topic modeling, and text clustering (Grubert & Siders, 2016) are promising because they offer ways to preserve "the superior abilities to interpret text holistically provided by humans but [incorporate] the formal rigor, reliability, and reproducibility of computer-assisted methods" (Nelson, 2020, p. 8).

Risk messages need to be precise and ecologically valid to the extent possible. Consequently, utilizing the language of the target population, as identified in source texts (e.g., interviews), in message construction is critical but also presents a substantive research challenge. At its heart, the operational challenge is to objectively identify, classify, and rank the importance of descriptive terms most strongly associated with each treatment condition from numerous source texts while also accounting for the variability in lengths of source texts. We confront this research challenge via judicious integration of NLP into qualitative research.

Detailed Procedure

Our procedure comprises four phases, two of which are qualitative and two of which are quantitative (Table 2). Integration occurred as the products from one phase provided critical inputs for the subsequent phases (Figure 1). The cumulative effect of each instance of integration enabled us to utilize locally derived language in our final messages with precision.

Qualitative Phase I

The procedure begins with Qualitative Phase I, comprising Step 1 Qualitative source text and Step 2 Human coding. Briefly, in Step 1, we compiled our source text by conducting and subsequently transcribing semi-structured interviews with 45 individuals in three flood-prone communities in our study area (Figure 2). In Step 2, we used human coding to bin local language from the semi-structured interviews into character language categories (hero vs. victim) based on a NPF codebook (Shanahan et al., 2018a). Integration occurred as the human-coded hero and victim texts became the foundation of the subsequent quantitative phase. Below, we detail each Step outlined in Figure 1 and Table 2.

Table 2. Purpose of Each Step in the Procedure.

			Contributions to Message Language Precision				
Phase	Procedure Step	Brief Description	Theoretical Foundation	Ecological Validity	Construct Validity	Internal Validity	Reliability
Qualitative I	1. Qualitative source text	Collect target audience language via semi-structured interviews	Raw material for operationalizing NPF concepts	Uses real-world language from audience	Audience contributes to measurement		
	2. Human coding	Categorize language from qualitative source text	Application of NPF concepts and categories	Uses real-world language from audience		Improves measurement for causal tests	Inter-coder reliability assessed
Quantitative I	3. Natural language processing	Parse source text using computational approaches relevant for developing message treatments	Objectively discern word sets to precisely operationalize NPF mechanism of character		Filters relevant information	Improves measurement for causal tests	
	4. Word classification	Provide relative ranking of words within each corpus	Identification of specific words to operationalize NPF concepts	Uses real-world language from audience	Assigns words distinctly to concepts	Improves measurement for causal tests	
Qualitative II	5. Algorithmic message construction	Formulate messages with parallel structure	Message differentiation based on NPF concepts	Uses real-world language from audience	Discriminates clearly among treatments	Improves measurement for causal tests	Parallel structure allows for testing reliability
Quantitative II	6. Validity & reliability testing	Evaluates persuasiveness of messages	Expectations are based on NPF	Tests conducted with real-world audience	Different treatments should produce different results	Establishes measurement properties for causal tests	Repeated testing of replicated language

Step 1 Qualitative Source Text. We conducted semi-structured interviews to provide the vernacular needed to build narratives in the target audience's own language; this step was conducted to improve the operationalization of NPF theory and to reduce threats to ecological and construct validity (Table 2). Thus, the raw material for narrative construction came from semi-structured interviews conducted with 45 individuals in three communities along the Yellowstone River in Montana. These three communities–Livingston, Miles City, and Glendive–were chosen because they border the Yellowstone River and had all experienced significant riverine floods recently despite manmade levees intended to protect infrastructure. Regardless of these commonalities, these communities have different relationships with the river, including varying recreational and economic opportunities (Shanahan et al., 2018c; Bergmann et al., 2020). The purposive sampling procedure aimed to achieve a sample with individuals from a range of affected sectors in the communities. The resulting sample included interested citizens, business owners, and residents from along the river. The interviews were distributed across the three communities ($n_{Livingston} = 11$, $n_{MilesCity} = 18$, $n_{Glendive} = 15$). We conducted these interviews from February-June 2017.

The first section of the interview protocol (Shanahan et al., 2019b) focused on problems, benefits, and risks associated with flooding on the Yellowstone River, as well as sources of information for learning about such flooding. To develop our message treatments, we needed locally derived language describing victim and hero characters. Thus, the second section asked about *harm* from flooding to elicit victim language and *preparation for and recovery from* flooding to elicit hero language.

The Human Ecology Learning and Problem Solving (HELPS) Lab at Montana State University transcribed nearly all the audio files from the interviews, with researchers completing the remaining few. In total, the 45 interviews resulted in 42 transcripts. Two individuals were interviewed simultaneously. Another two individuals refused audio recording per the informed consent procedures; field notes for these interviews were taken but not used subsequently. We aimed to allow interviews to unfold at a relatively leisurely pace so that interviewees would feel comfortable and would use their own descriptive language. The resulting transcripts ranging in length from about 3,500 words to over 32,000 words, with a median of 9,016 words.

Step 2 Human Coding. We used human coding to assign local language from the 42 semi-structured interviews (Step 1 Qualitative source text) into appropriate narrative elements (e.g., characters) and nodes within those elements (e.g., hero or victim language). More plainly, we manually tagged victim and hero language in all interview transcripts based on NPF theory. This step aimed to fortify the integration of theory into final message construction while simultaneously bolstering the ecological and internal validity of final message treatments (Table 2).

Human coding for characters was an iterative process that began in a deductive manner. Previous NPF codebooks (Shanahan et al., 2013) provided the foundation for the coding. Existing NPF research also provided definitions for the character nodes. According to the NPF, heroes are fixers of problems, whereas victims are entities being harmed (Shanahan et al., 2018b). Four researchers began by independently coding the same transcript in NVivo11 software (QSR International Pty Ltd., 2015). The main nodes, established deductively from the NPF, were the hero and victim character categories. The specific identities of these characters (i.e., the sub-nodes) emerged inductively from the data (e.g., government floodplain administrator under the hero node or individual homeowner under the victim node). The researchers then convened to compare specific coding actions and categories. Based on this comparison, they revised and consolidated the codebook. Three of these researchers then independently coded a second transcript in full. They met again to refine the node structure and coding scheme. These iterative comparisons were important for ensuring reliability in coding. The researchers then distributed and coded the remaining 40 transcripts based on the refined coding scheme, coding at the sentence level for hero

and victim language. A fourth coder subsequently coded a random selection of 20% of each interview to check for inter-coder reliability. Averaged across all interviews, Cohen's kappa (Cohen, 1960) for hero coding was 0.883 and for victim coding was 0.880, which indicates substantial agreement (Landis & Koch, 1977).

Quantitative Phase I

Quantitative Phase I comprises Step 3 Natural Language Processing and Step 4 Word classification. In this phase, we employed NLP techniques and word classification to distinguish words from interview transcripts that were most strongly associated with each of our treatment conditions. Integration occurred as the individual "hero words" and "victim words," identified via NLP and word classification, provided our research team with the key terms to use in the narratives to precisely operationalize hero versus victim message treatment conditions.

Step 3 Natural Language Processing. Across all interviews (Step 1 Qualitative source text), the human-coded text associated with characters hero and victim characters (identified in Step 2 Human coding) was combined into bodies (i.e., corpora) of character-related text: one corpus for hero language and one for victim. In turn, these corpora were subjected to NLP to identify and rank word choices for each character type. The rationale for integrating computational techniques is twofold. First, we wanted to reduce threats to the internal and construct validity of our final messages (Table 2) by efficiently and objectively discerning the words that most precisely characterized victim or hero treatments to the target audience. Second, the corpora of character-related text were large and unwieldy. Specifically, the hero corpus contained about 35,400 words, while the victim corpus contained about 58,300 words. In what follows, each step used in our computational NLP approach is described.

Assessment of the coded text using NLP techniques required carrying out certain preprocessing procedures. Natural language (i.e., human-generated language) presents a combinatorial problem for computers, which can "view" each unique letter, word, sentence, and paragraph as a feature for consideration. This high dimensionality can dramatically slow down automated content analysis algorithms. Thus, the goals of preprocessing are to reduce the number of features in a narrative without losing relevant information and to reach a vectorized representation for computational text analysis models. All preprocessing steps used the RStudio integrated development environment (RStudio Team, 2019) and the R programming language (R Core Team, 2019), relying heavily on the tm (text mining) package (Meyer et al., 2008).

First, we reorganized the 42 coded, semi-structured interview transcripts (Step 1 Qualitative source text and Step 2 Human coding) into sets of documents by label (i.e., hero and victim) so that each document contained all the coded language from a label found in an interview. For example, document 1 in the hero corpus contained all the hero-coded language elements from interview 1, whereas document 1 in the victim corpus contained all the victim-coded language elements from the same interview. In total, we extracted 472 instances of hero language elements and 748 instances of victim language elements. These language elements ranged from one sentence to a paragraph in length. The aggregation of each set of documents made up the hero and victim corpora, respectively.

The next set of four preprocessing steps included commonly used approaches in automated content analysis: conversion to lowercase, character scrubbing, stop-word removal, and tokenization. All of these methods reduce the number of features for consideration with minimal semantic loss. Lowercase conversion quickly reduces the number of features considered, as words like "he" and "He" would otherwise be interpreted as unique terms. Alphanumeric character (i.e., letter) scrubbing removes unhelpful symbols, such as punctuation, URL markers, and numbers.

Similarly, stop-word removal eliminates many high frequency terms used in natural language such as "a," "the," and "that." The tm R-package default list of 174 English stop-words and an additional custom list, created by our researchers, were used for selecting words tagged for removal from the documents. The custom list was tailored to reflect interview transcripts and the flood risk domain. This list included terms like "uh," "uhm," "hmm," which are important social cues in vocal speech but not relevant to the formation of narratives.

The final preprocessing step, tokenization, breaks the documents into feature vectors. These vectors are sequences of integers that store the counts for each unique term in every document of the corpus; each integer in a vector represents a count for a term from a document. In order to tokenize, a term length must be determined. For this project, the documents were broken into unigram terms (i.e., one word per term). Bigram (i.e., two words per term) and N-gram (i.e., n words per term) models were explored, but they did not yield useful information. The feature vectors are combined into a term-document matrix, where rows represent the unique terms found in the corpus, columns represent the documents of the corpus, and cells store the term count. This creates a large and sparse matrix from which we can perform automated content analysis.

The performance of algorithms that use a vector approach for storing the word frequencies is linear (i.e., $O(n)$) and was entirely sufficient for our purposes. Run time on a commercially available laptop was a nonissue for our dataset, and optimizing computational performance was not a goal nor necessary in this step. Rather, our focus was on using the NLP results in the greater mixed methods procedure. However, if our dataset had been large enough to warrant faster computational search times, we would have compared the vector storage approach with alternative data structures (e.g., tree) to determine which storage approach optimized computational performance.

Step 4 Word Classification. We classified the words in the hero and victim corpora (Step 3 Natural Language Processing) using automated content analysis. The purpose of this step was to classify the "hero words" and "victim words" to operationalize NPF theory most precisely in the final messages while also reducing threats to ecological, construct, and internal validity (Table 2). We experimented with four different content analysis techniques and found that term frequency calculations proved to be the most informative text analysis techniques for the creation of narratives (see King (2019) for full description of the other three methods). Term frequency measurements on transcripts from the target audience provided the exact vocabulary used to communicate messages about the flood domain. Using the term-document matrices, the term counts for each corpus were calculated by summing across each row. Given that the corpora were of different sizes, term counts were normalized by dividing by the total number of words in each corpus, calculating a relative frequency, Rf. Some terms appeared frequently in both corpora. Using a difference of proportions method (King, 2019; Shanahan et al., 2019), we subtracted the relative frequency of a term for one corpus, Rf_x from its relative frequency for another, Rf_y. The difference of proportions allowed for the ranking of words by their relative importance to each corpus. Given that narratives were to focus on hero and victim characters, we looked specifically at the hero minus victim proportions, $Rf_H - Rf_V$. Terms with large positive values were hero terms; terms with large negative values were victim terms. We took the head and tail of this spectrum (the top and bottom 4% of terms) to create the hero and victim vocabularies that informed the eventual narrative construction.

Qualitative Phase II

Qualitative Phase II comprises only one step, Step 5 Algorithmic message construction. Here, we employed a human-generated algorithm—rooted in narrative theory—to construct the narratives using key words discovered through the NLP analysis. Integration occurred again between this phase and the subsequent one, as the algorithmic message construction enabled us to evaluate the instrument fidelity of each segment of each message treatment in Quantitative Phase II.

Step 5 Algorithmic Message Construction. With the hero and victim vocabularies in hand (Step 4 Word classification), we proceeded with algorithmic message construction. Algorithmic message construction reduced threats to reliability and ecological, construct, and internal validity (Table 2). As discussed earlier, the primary goal in the exemplar was to investigate the influence of the narrative mechanism of victim and hero characters on affective responses (Shanahan et al., 2019) and intended risk mitigation behaviors (Raile et al., 2022). As such, we constructed narrative messages with language corresponding to three distinct character mechanisms—victim, hero, and victim-turns-hero. Victim language emphasizes negative outcomes for the audience members and their communities. Hero language emphasizes the entities responsible for fixing flood-related problems, including the audience members. Victim-to-hero language creates an arc in which the negative outcomes can be overcome by the audience members and their communities.

The secondary goal in the exemplar was to determine whether science information presented in the language of probability or certainty had greater persuasive power. Probability language is the *de facto* standard for describing riverine flood risk in the United States, whereas certainty language is emerging as a new approach for describing earthquake risks (Jones, 2019). Thus, the science information in each message was described with either probability or certainty language.

We constructed each narrative with a common structure; however, we strategically varied the content of each of the four segments (or pieces) that compose a message to enable testing of different treatment combinations (Figure 3; Shanahan et al., 2019). All narrative messages opened with an identical definition of a riverine flood. The second segment in each narrative framed the problem of flooding with either a victim, hero, or victim-turns-hero frame. The third segment described science information about flooding using either probability or certainty language. The fourth and final segment described how the characters in the story took action to prepare for a flood hazard with a character mechanism of victim, hero, or victim-turns-hero. Thus, narrative messages for the victim treatment included victim language in both the second (problem framing) and fourth (characters in action) segments of the messages; likewise, narrative messages for the hero treatment included hero language in both of these segments. In contrast, the narrative messages for the victim-turns-hero treatment include a combination of victim and hero language. The full narrative messages with segments identified are presented in S1 Text of Shanahan et al. (2019).

To improve internal validity, construct validity, and reliability in message construction, we constructed a "word use signature" histogram for each narrative message by plotting the frequency of $Rf_H - Rf_V$ scores (zero indicating neutral between hero and victim) for the words in each narrative (Figure 4). A word that appears more than once in a message also shows up more than once in the histogram. As anticipated, the victim treatments are skewed left, hero treatments are skewed right, and victim-turns-hero treatments are centered closer to zero. Notably, it was not possible to construct a hero narrative with only words with positive $Rf_H - Rf_V$ scores or a victim narrative with only negative $Rf_H - Rf_V$ scores. For instance, it was imperative that we used the words "river" and "flood" in all treatments, and these particular words had $Rf_H - Rf_V$ scores of +14.9 and −12.6, respectively.

Quantitative Phase II

The final phase of our research is Quantitative Phase II. This phase comprises one step, Step 6 Validity & reliability testing, wherein we evaluated the instrument fidelity of the message treatment conditions by conducting validity and reliability testing. To do so, we returned to the three flood-prone communities (Figure 2) and asked 90 participants to evaluate each narrative message using dial response testing.

Step 6 Validity and Reliability Testing. Step 6 reduced threats to reliability and ecological, construct, and internal validity (Table 2). The exemplar's full experimental protocol and results and our

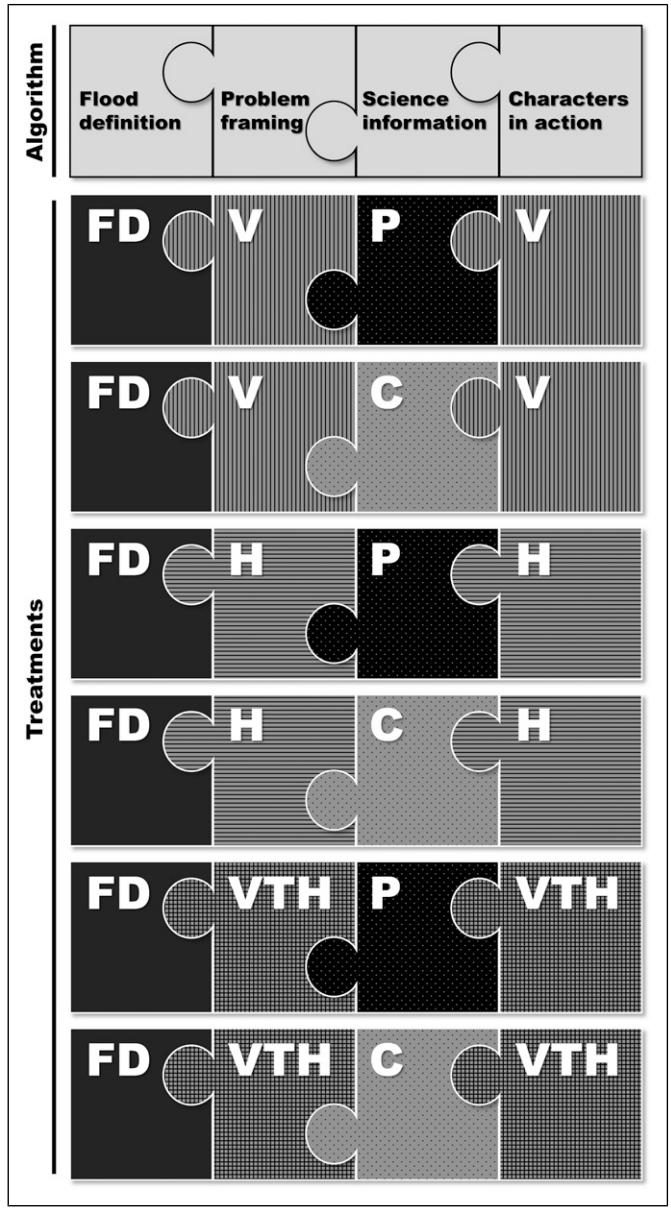

Figure 3. Algorithmic message construction. Each message treatment (row) is comprised of four segments, which can be conceptualized as four distinct puzzle pieces (columns). Variation in the composition of the segments results in distinct message structure. The flood definition [FD] is identical across treatments. The problem framing and characters in action segments are each comprised of one of three sets of text corresponding to character treatment: victim [V], hero [H], and victim-turns-hero [VTH]. The science information segment contains information about riverine flood risk described in either a probability [P] or certainty [C] context. Where labels and fill are identical within columns, the language in those segments of the narratives is identical. For instance, the victim problem framing language is identical in the first and second treatment rows.

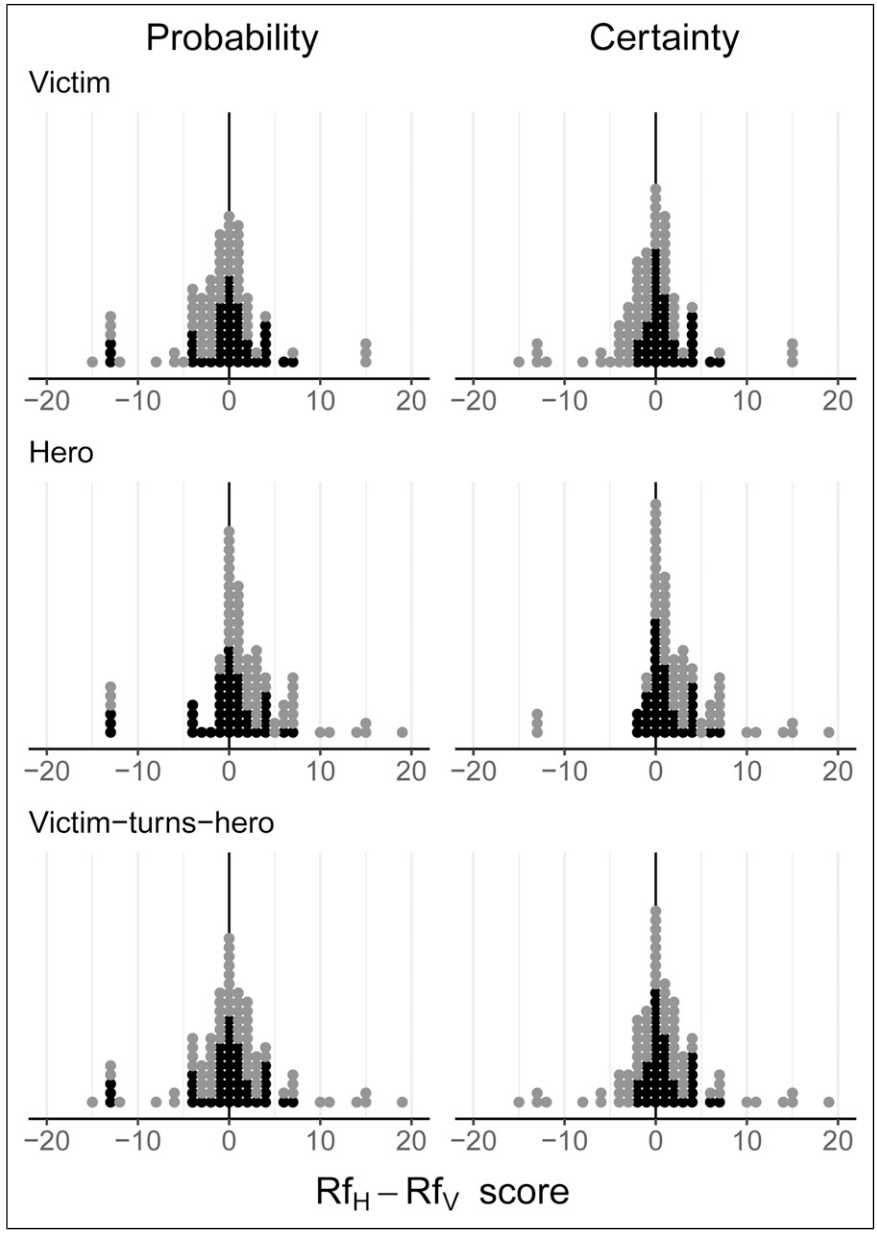

Figure 4. Histograms of $Rf_H - Rf_V$ scores for words used in the victim, hero, and victim-turns-hero narrative treatments associated with either probability or certainty to describe the embedded science language. Each dot represents one word. Black dots represent words from the "flood definition" and "science information" sections of each narrative (Figure 3), which are constant within column. Gray dots represent words from the "problem framing" and "characters in action" sections, which vary with character treatment. Victim treatments are skewed left, hero treatments are skewed right, and victim-turns-hero treatments are centered near zero.

interpretation of the validity and reliability testing are published in Shanahan et al. (2019). Briefly, the three communities that were the sites of the semi-structured interviews also became the sites for field testing of the eight risk communication messages. The goal, again, of the exemplar was to

test the language with audiences from the same places as the individuals who generated the vocabularies via the semi-structured interviews (Shanahan et al., 2019). The testing technology required the construction of videos with audio for all messages. The videos were recorded using Microsoft PowerPoint with white words on dark blue backgrounds and audio overlays. Each slide contained a single sentence from the message to prevent audience members from reading ahead. The narrator attempted to remain as calm and impassive as possible when reading the messages to focus audience members on the content alone.

To obtain a sample of participants to test these eight messages, the researchers ordered a random sample of 500 addresses from Survey Sampling International for each of the three study communities. Postcards went out to these addresses inviting one adult from the household to participate and offering a $50 incentive in return. The research sessions took place in the respective communities on prearranged dates in October and November of 2017. Potential participants could sign up via the website of the HELPS Lab. A second postcard went out to nonrespondents 2 weeks later and invited individuals to spread news about the sessions. We also advertised via local newspapers and social media accounts linked to city governments. Our research team conducted four sessions in each community. The final sample included 90 research participants: 36 from Livingston, 22 from Miles City, and 32 from Glendive. We held multiple sessions in each community, with the number of participants ranging from 4 to 11 in each session. The final sample was nearly evenly split in terms of women and men but did skew somewhat older than the general populations of adults in these communities.

The test sessions, which lasted approximately 1 hour each, featured dial response technology and a follow-up focus group and demographic survey. The dial response was used to measure affective response, a dimension of narrative transportation that measures audience engagement (Green & Brock, 2000). The dial response technology, the Perception Analyzer™ from Dialsmith, permits instantaneous and continuous measurement of audience response to either live or recorded messages. Participants hold dials with preloaded data ranges as specified in the software. For this study, response options ranged from 0–100. The middle (vertical) position of the dial indicated 50 and was the neutral score. Participants were instructed to respond throughout the message with regard to how positive or negative the message was making them feel (i.e., their affective response to the message). The facilitator asked participants to start at the neutral position of 50 and indicated that 0 was the most negative score and 100 was the most positive score. Each session included a brief practice with using the dial response technology. The researchers randomized the order of the eight risk communication messages across sessions to eliminate message order effects. The software recorded each participant dial once per second.

The results from these sessions were used to test hypotheses about affective responses to character language in narrative science messages and to the type of science language (probability vs. certainty) that described flood hazard risk as part of the persuasion process (Shanahan et al., 2019). From this testing, we learned that participant responses differed among message treatments. Altering the narrative mechanism of character selection in messages consistently resulted in differences in participant responses; participants had slightly negative responses to victim treatments but positive responses to hero and victim-turns-hero treatments. These results largely corresponded with our predictions, thereby suggesting that we had minimized threats to construct validity. In simple terms, we had measured the concepts (hero and victim characters) that we had intended to measure based on theory. Such construct validity would be crucial to internal validity (i.e., establishment of cause and effect) in our later experiment. We found no differences between the probability and certainty versions of the science statements, which both produced negative affective responses across treatments. However, we did find remarkably consistent aggregate responses to the flood definition and science information segments, which provided evidence of reliability in the measurement. In sum, we concluded that our process minimized threats to validity

and reliability. Had this testing revealed problems, we would have returned to message construction to evaluate which step might have been problematic.

Having determined that the narrative messages satisfy construct validity and precisely operationalize the treatment conditions, we used them in a mail survey of residents who live along the Yellowstone River, to test whether different narrative science messages have differential effects on affective response and intended risk preparation behavior (Raile et al., 2022). Figure 5 presents a research process display of our sequential mixed methods procedure, providing details of the optimization of instrument fidelity and details of what Onwuegbuzie et al. (2010, p. 58) refer to as crossover analyses, "which involves using one or more analysis types associated with one tradition (e.g., quantitative analyses) to analyze data associated with a different tradition (e.g., qualitative data)."

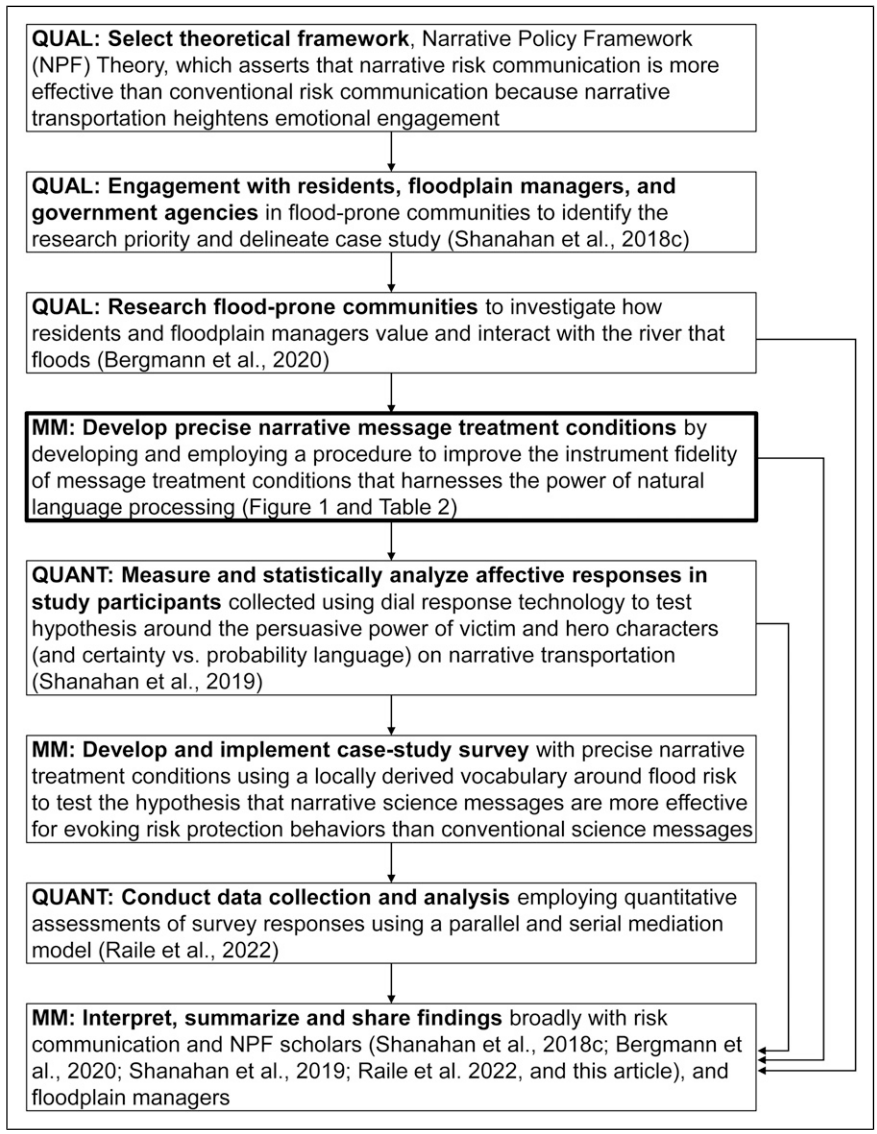

Figure 5. Research process display. How the exploratory sequential mixed methods procedure (bolded box) is embedded into the broader research investigating riverine-flood risk communication. MM: mixed methods; QUAL: qualitative; QUANT: quantitative.

Discussion

Contribution to the Field of Mixed Methods Research

Our procedure makes a unique contribution to the field of mixed methods research in two ways as we address Onwuegbuzie et al.'s call (2010) for "more publications…that outline explicitly ways of optimizing the development of instruments by mixing qualitative and quantitative techniques" (pp. 57–58). First, we address the black box of experimental message treatment construction with what Onwuegbuzie et al. (2010) refer to as crossover analysis. We do so by blending a constructivist stance (i.e., perspective of residents through interviews, human coding, and message construction) with a positivist stance (i.e., NLP, word classification, and validity and reliability testing). Additionally, we employ a compatible theoretical foundation in the NPF that brings an objective epistemological approach (i.e., objective measures of universal narrative structure such as characters) to bear on a subjective ontology (i.e., social construction of reality through narratives). Second, we offer guidance on a powerful and relatively new tool in textual analysis, that of NLP, for use in mixed methods research. This use of NLP in crossover analyses between inductive and deductive logics optimizes instrument fidelity by linking theory (i.e., NPF) with qualitative data (i.e., interviews, coding, and message construction) and quantitative data (i.e., words identified via NLP). In turn, we sought to validate our message construction through further quantitative measure, that of affective response to different message treatments.

The novelty of this study is harnessing the power of integration through crossover analysis to develop a mixed methods approach to improve instrument fidelity in message testing research by addressing the critical need of developing a procedure for precisely constructing message treatments via broadening the use of NLP in the social sciences. Integration is a challenge in mixed methods research (Bryman, 2007; Fetters et al., 2013; Uprichard & Dawney, 2016) but is important to surmount because integration "produces a sum greater than the individual parts" (Fetters & Freshwater, 2015, p. 208). In particular, our work highlights how integration in the Research design dimension improves the Research integrity dimension, that is, precision in message construction (Table 2; see also section Detailed Procedure) (Fetters & Molina-Azorin, 2017). Our procedure for constructing precise messages is a research outcome that resulted from integration in the following dimensions: Rationale; Study purpose, aims, and research questions; Researcher; Team; Data collection; Data analysis; and Interpretation (Fetters & Molina-Azorin, 2017).

The integration in the Rationale and Study purpose, aims, and research questions dimensions emerged from a clear need to open the black box of developing message treatments to overcome the numerous potential threats to validity and reliability that arise from depending on expert opinion to construct treatments. In the Researcher dimension, our research team (i.e., the co-authors on this article) were drawn together to address the challenge of constructing precise message treatments because of experiences that lead each to highly value employing mixed methods procedures; without question, integration in this dimension was the bedrock upon which the integration in all other dimensions was built. For instance, in the Team dimension, each researcher was a domain expert in fields ranging from political science, human geography, economics, hydrology, to computer science. This diversity in team expertise brought incredible creativity and energy but also many challenges. As others have noted (Poth, 2019), our team quickly learned that integration is hard work. Each team member was stretched to learn the key concepts, theory, history, and vernacular of the other disciplines as related to the common research goal and to communicate the nuance and importance in their own discipline using a common language. As a result, frequent meetings were required wherein patience, humility, humor, and

excellent team leadership were critically important to advancing the research. Perhaps not surprisingly, some of the most productive and lively conversations arose as the team carefully considered the strengths of existing qualitative and quantitative approaches to select the best mixed method approaches to integrating NLP with NPF in the Data collection, Data analysis, and Interpretation dimensions.

To our knowledge, our team is the first to utilize NLP to enhance message treatment construction. Our efforts were not without challenges. We faced similar challenges in the Data collection, Data analysis, and Interpretation dimensions as those presented by a multitude of other scholars (Guetterman et al., 2018; Nelson, 2020; O'Halloran et al., 2018; Rohrer et al., 2017). For instance, as we strove to incorporate the most useful NLP techniques into our procedure, we explored several "dead ends" that we originally thought would be quite useful. In addition to the term frequency approach to word classification that we describe in the detailed procedure above, we also attempted three other approaches to identifying words associated with victim and hero characters. These approaches were topic modeling, sentiment analysis, and a formal classification algorithm (King, 2019). Topic modeling (Blei, 2012) refers to the application of quantitative techniques used to find common or unifying themes in a set of documents within a corpus. These techniques can be used to either confirm the existence of known topics within a corpus or to find latent topics not readily apparent–even to a trained domain expert. Sentiment analysis techniques (Cambria et al., 2013; Mäntylä et al., 2018) aim to measure emotions embedded in narratives. Two approaches to measuring sentiment are commonly used. The first approach is a nominal technique that classifies words into bins, where each bin can represent a sentiment (e.g., happy, sad, angry, etc.). The second technique uses ratio-scale measurements to calculate a polarity score for each word. Many techniques exist to assign and adjust polarity of words depending on context. Finally, classification algorithms (Han et al., 2009) are machine-learning based approaches that aim to reduce manual coding of information. By training a model with known data, classification algorithms can then be exercised with new, previously unseen data with the expectation that the model will yield the correct classification.

Briefly, the NLP methods of topic modeling and sentiment analyses generally confirmed researcher interpretation of the linkages amongst words in the corpora but did not provide new or unappreciated information to the research team. The formal text classification algorithm rendered only minimally useful information because the quantifiable aspects of victim and hero language—term frequencies—were higher in the victim documents simply because the victim documents were generally longer than the hero documents. Consequently, the classifier produced skewed results: precision and recall were moderate for the hero corpus (50–70%) but low for the victim corpus (<30%; King, 2019). Despite their limitations, each of these methods helped the research team better understand the corpora. However, only through a combination of intramethod analytics and core integration analytics was our team able to fully appreciate the strengths of the term frequency approach we ultimately employed. In the end, we agree that NLP is most useful when augmented by qualitative analysis (Guetterman et al., 2018) and that NLP offers improved integration of qualitative data into an exploratory sequential mixed methods research design (O'Halloran et al., 2018).

Considerations, Limitations, and Future Directions

The procedure presented here moves the theoretical discussions of mixed methods standards of validity and integration (Dellinger & Leech, 2007; Fàbregues & Molina-Azorín, 2017; Fetters et al., 2013) into practice. However, a reader might ask whether our approach was worth the considerable effort. Much of our effort was the result of exploring and comparing specific

methods, which might not be necessary in subsequent studies. Ultimately, our approach boils down to semi-structured interviewing, human coding, the production of relative word frequencies and their systemic application in message construction, and then some form of testing the validity and reliability properties of the resulting messages. The multiple stages and mixed methods necessitate a team approach, but no single piece is exceedingly difficult on its own. At this point, the validity and reliability in other approaches remain unknown, so comparing the ratio of labor to precision is impossible. However, moving forward, researchers can be more intentional in evaluating this ratio. Thus, the primary limitation of our research is that we cannot explicitly state if our procedure is "worth it" for other researchers or "how much better" our procedure is over black-box message construction.

Our future research directions seek to transport our mixed methods process to other domains such as viral spillover (e.g., coronavirus, Ebola) and cyber security. Indeed, the accuracy of risk communication studies in these domains has the potential to save lives and increase security at multiple levels—personal, municipal, state, national. Applications across different field domains will also test the transportability of our mixed methods approach.

Conclusions

Our procedure improves instrument fidelity in message testing research via a novel integration of qualitative and quantitative methods to address a critical research need: bolstering theoretical grounding, validity, and reliability as forms of message precision. We found this procedure to be effective for our purposes and suspect it will prove useful beyond our research domains of narrative communication and hazard preparedness. Our research team looks forward to its use and improvement in future studies.

Acknowledgments

This research team is grateful for each of the people who we interviewed in the field and who participated in the validity and reliability testing. We thank the many floodplain managers who generously shared their time and knowledge with us, especially during the initial phases of this research. We also thank Kate French for her insights during the early phases of this research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: All authors of this study received financial support from National Science Foundation (Grant Number 1635885). In addition, AMR received partial support from the National Science Foundation EPSCoR Cooperative Agreement OIA-1757351; JM and RCR were funded through National Science Foundation RII Track-1 with the Montana Institute on Ecosystems (EPS-1101342 and OIA-1443108); and GCP was supported by the USDA National Institute of Food and Agriculture (Hatch Project Number 1015745). The National Institute of General Medical Sciences of the National Institutes of Health also provided support through their funding for Montana INBRE and for the Human Ecology Learning and Problem Solving Lab at Montana State University – Bozeman (Award Number P20GM103474).

ORCID iD

Ann Marie Reinhold https://orcid.org/0000-0003-0411-3486

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-t

Barbour, J. B., Doshi, M. J., & Hernández, L. H. (2015). Telling global public health stories: Narrative message design for issues management. *Communication Research*, *43*(6), 810–843. https://doi.org/10.1177/0093650215579224

Barbour, J. B., Doshi, M. J., & Hernández, L. H. (2016). Telling global public health stories: Narrative message design for issues management. *Communication Research*, *43*(6), 810–843. https://doi.org/10.1177/0093650215579224

Barnes, P. H. (2002). Approaches to community safety: Risk perception and social meaning. *Australian Journal of Emergency Management*, *1*(1), 15–23.

Bergmann, N. T., McEvoy, J., Shanahan, E. A., Raile, E. D., Reinhold, A. M., Poole, G. C., & Izurieta, C. (2020). Thinking Through Levees: How Political Agency Extends Beyond the Human Mind. *Annals of the American Association of Geographers*, *110*(3), 827–846. doi: 10.1080/24694452.2019.1655387.

Blei, D. M. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities*, *2*(1), 8–11.

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research*, *1*(1), 8–22. https://doi.org/10.1177/2345678906290531

Bryman, A. (2016). *Social research methods*. Oxford University Press.

Bubeck, P., Botzen, W. J. W., & Aerts, J. C. J. H. (2012). A review of risk perceptions and other factors that influence flood mitigation behavior. *Risk Analysis*, *32*(9), 1481–1495. https://doi.org/10.1111/j.1539-6924.2011.01783.x

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15–21. https://doi.org/10.1109/mis.2013.30

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Collins, K. M., Onwuegbuzie, A. J., & Sutton, I. L. (2006). A model incorporating the rationale and purpose for conducting mixed methods research in special education and beyond. *Learning Disabilities: A Contemporary Journal*, *4*(1), 67–100.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Dellinger, A. B., & Leech, N. L. (2007). Toward a unified validation framework in mixed methods research. *Journal of Mixed Methods Research*, *1*(4), 309–332. https://doi.org/10.1177/1558689807306147

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and perspectives*, *38*(1), 105–123.

Fàbregues, S., & Molina-Azorín, J. F. (2017). Addressing quality in mixed methods research: A review and recommendations for a future agenda. *Quality & Quantity*, *51*(6), 2847–2863. https://doi.org/10.1007/s11135-016-0449-4

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—principles and practices. *Health Services Research*, *48*(6 Pt 2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

Fetters, M. D., & Freshwater, D. (2015). *Publishing a methodological mixed methods research article*. Thousand Oaks, CA: Sage Publications.

Fetters, M. D., & Molina-Azorin, J. F. (2017). *The Journal of Mixed Methods Research starts a new decade: The mixed methods research integration trilogy and its dimensions*. Thousand Oaks, CA: Sage Publications.

Fishbein, M. (2008). A reasoned action approach to health promotion. *Medical Decision Making*, *28*(6), 834–844. https://doi.org/10.1177/0272989X08326092

Flanders, J. (2005) Detailism, digital texts, and the problem of pedantry. *Text Technology*, *14*(2), 41.

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, *79*(5), 701–721. https://doi.org/10.1037//0022-3514.79.5.701

Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS, Political Science & Politics*, *48*(1), 80. https://doi.org/10.1017/s1049096514001784.

Grubert, E., & Siders, A. (2016). Benefits and applications of interdisciplinary digital tools for environmental meta-reviews and analyses. *Environmental Research Letters*, *11*(9), 093001. https://doi.org/10.1088/1748-9326/11/9/093001

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research*, *20*(6), Article e231. https://doi.org/10.2196/jmir.9702

Han, H.-q., Zhu, D.-H., & Wang, X.-f. (2009). Semi-supervised text classification from unlabeled documents using class associated words. In Paper presented at the 2009 International Conference on Computers & Industrial Engineering, France, 6-9 July 2009. https://doi.org/10.1109/iccie.2009.5223918

Hennink-Kaminski, H. J., & Dougall, E. K. (2009). Tailoring hospital education materials for the period of purple crying: Keeping babies safe in North Carolina media campaign. *Social Marketing Quarterly*, *15*(4), 49–64. https://doi.org/10.1080/15245000903348772

Howell Smith, M. C., Babchuk, W. A., Stevens, J., Garrett, A. L., Wang, S. C., & Guetterman, T. C. (2020). Modeling the use of mixed methods–grounded theory: Developing scales for a new measurement model. *Journal of Mixed Methods Research*, *14*(2), 184–206. https://doi.org/10.1177/1558689819872599

Jones, L. (2019). *The big ones: How natural disasters have shaped us (and what we can do about them)*. Anchor.

Jordan, A., Taylor Piotrowski, J., Bleakley, A., & Mallya, G. (2012). Developing media interventions to reduce household sugar-sweetened beverage consumption. *The Annals of the American Academy of Political and Social Science*, *640*(1), 118–135. https://doi.org/10.1177/0002716211425656

Khanal, R. C. (2013). Concerns and challenges of data integration from objective post-positivist approach and a subjective non-positivist interpretive approach and their validity/credibility issues. *Journal of the Institute of Engineering*, *9*(1), 115–129. https://doi.org/10.3126/jie.v9i1.10677

King, H. (2019). *Informing the Construction of Narrative-based Risk Communication*. [Master's thesis, Montana State University]. ScholarWorks. https://scholarworks.montana.edu/xmlui/bitstream/handle/1/15773/king-informing-the-2019.pdf?sequence=3

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lapka, C., Jupka, K., Wray, R. J., & Jacobsen, H. (2008). Applying cognitive response testing in message development and pre-testing. *Health Education Research*, *23*(3), 467–476. https://doi.org/10.1093/her/cym089

Leshner, G., Bolls, P., Gardner, E., Moore, J., & Kreuter, M. (2018). Breast cancer survivor testimonies: Effects of narrative and emotional valence on affect and cognition. *Cogent Social Sciences*, *4*(1), 1426281. https://doi.org/10.1080/23311886.2018.1426281

Ludy, J., & Kondolf, G. M. (2012). Flood risk perception in lands "protected" by 100-year levees. *Natural Hazards*, *61*(2), 829–842. https://doi.org/10.1007/s11069-011-0072-6

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002

McLaughlin, B. (2020). Tales of conflict: narrative immersion and political aggression in the United States. *Media Psychology*, *23*(4), 579–602. https://doi.org/10.1080/15213269.2019.1611452

McLaughlin, B., Velez, J. A., Gotlieb, M. R., Thompson, B. A., & Krause-McCord, A. (2019). React to the future: political visualization, emotional reactions and political behavior. *International Journal of Advertising*, *38*(5), 760–775. https://doi.org/10.1080/02650487.2018.1556193

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54. https://doi.org/10.18637/jss.v025.i05

National Cancer Institute. (2005). Theory at a glance: A guide for health promotion practice. A publication of the U.S. Department of Health and Human Services and National Institutes of Health.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Occa, A., & Suggs, L. S. (2016). Communicating breast cancer screening with young women: An experimental test of didactic and narrative messages using video and infographics. *Journal of Health Communication*, *21*(1), 1–11. https://doi.org/10.1080/10810730.2015.1018611

O'Halloran, K. L., Tan, S., Pham, D.-S., Bateman, J., & Vande Moere, A. (2018). A digital mixed methods research design: Integrating multimodal analysis with data mining and information visualization for big data analytics. *Journal of Mixed Methods Research*, *12*(1), 11–30. https://doi.org/10.1177/1558689816651015

Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, *4*(1), 56–78. https://doi.org/10.1177/1558689809355805

Poehlman, J. A., Sidibe, T., Jimenez-Magdaleno, K. V., Vazquez, N., Ray, S. E., Mitchell, E. W., & Squiers, L. (2019). Developing and testing the Detén El Zika campaign in Puerto Rico. *Journal of Health Communication*, *24*(12), 900–911. https://doi.org/10.1080/10810730.2019.1683655

Poth, C. (2019). Realizing the integrative capacity of educational mixed methods research teams: Using a complexity-sensitive strategy to boost innovation. *International Journal of Research & Method in Education*, *42*(3), 252–266. https://doi.org/10.1080/1743727x.2019.1590813

QSR International Pty Ltd. (2015). *NVivo (Version 11)*. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

R Core Team. (2019). *R: A language and environment for statistical computing*.

Raile, E. D., Shanahan, E. A., Ready, R. C., McEvoy, J., Izurieta, C., Reinhold, A. M., Poole, G. C., Bergmann, N. T., & King, H. (2022). Narrative Risk Communication as a Lingua Franca for Environmental Hazard Preparation. *Environmental Communication*, *16*(1), 108–124. doi: 10.1080/17524032.2021.1966818

Rohrer, J. M., Brümmer, M., Schmukle, S. C., Goebel, J., & Wagner, G. G. (2017). "What else are you worried about?"–Integrating textual responses into quantitative social science research. *PLoS ONE*, *12*(7), Article e0182156. https://doi.org/10.1371/journal.pone.0182156

RStudio Team. (2019). *RStudio*. Integrated Development for R. RStudio, Inc.

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy*, *2*(4), 419–436. https://doi.org/10.1207/S15327078IN0204_02

Semmler, S. M., & Loof, T. (2019). Audio-only character narration overcoming resistance to narrative persuasion. *Communication Research Reports*, *36*(3), 191–200. https://doi.org/10.1080/08824096.2019.1598855

Shanahan, E. A., Jones, M. D., McBeth, M. K., & Lane, R. R. (2013). An angel on the wind: How heroic policy narratives shape policy realities. *Policy Studies Journal*, *41*(3), 453–483. doi: 10.1111/psj.12025

Shanahan, E. A., Jones, M. D., & McBeth, M. K. (2018a). How to conduct a narrative policy framework study. *Social Science Journal 55*(3): 332–345. https://doi.org/10.1016/j.soscij.2017.12.002

Shanahan, E. A., Jones, M. D., McBeth, M. K., & Radaelli, C. (2018b). The narrative policy framework. In C. M. Weible & P. A. Sabatier (Eds.), *Theories of the policy process* (4th ed., pp. 173–213). Boulder, CO: Westview Press.

Shanahan, E. A., Raile, E. D., French, K. A., & McEvoy, J. (2018c). Bounded stories: How issue frames and narrative settings help to construct policy realities. *Policy Studies Journal 46*(4), 922–948. https://doi.org/10.1111/psj.12269

Shanahan, E. A., Adams, S. M., Jones, M. D., & McBeth, M. K. (2019a). The blame game: Narrative persuasiveness of the Intentional causal mechanism. In M. D. Jones, E. A. Shanahan, & M. K. McBeth (Eds.), *The Science of Stories: Applications of Narrative Policy Framework* (pp. 69–88). New York, New York: Palgrave Macmillan.

Shanahan, E. A., Reinhold, A. M., Raile, E. D., Poole, G. C., Ready, R. C., Izurieta, C., McEvoy, J., Bergmann, N. T., & King, H. (2019b) Characters matter: How narratives shape affective responses to risk communication. *PLoS ONE*, *14*(12): e0225968. doi: 10.1371/journal.pone.0225968

Uprichard, E., & Dawney, L. (2016). Data diffraction: Challenging data integration in mixed methods research. *Journal of Mixed Methods Research*, *13*(1), 19–32. https://doi.org/10.1177/1558689816674650

Wachinger, G., Renn, O., Begg, C., & Kuhlicke, C. (2013). The risk perception paradox—implications for governance and communication of natural hazards. *Risk Analysis*, *33*(6), 1049–1065. https://doi.org/10.1111/j.1539-6924.2012.01942.x

Whitlock, C., Cross, W., Maxwell, B., Silverman, N., & Wade, A. (2017). *Montana climate assessment*. Montana Institute on Ecosystems, Montana State University, and University of Montana.

Willoughby, J. F., & Brickman, J. (2020). Adding to the message testing tool belt: Assessing the feasibility and acceptability of an EMA-style, mobile approach to pretesting mHealth interventions. *Health communication 36*(10), 1260–1267. https://doi.org/10.1080/10410236.2020.1750748