Estimating Climate Change Numbers:

Mental Computation Strategies That Can Support Science Learning

Ian Thacker

Hannah French

University of Texas, San Antonio

Shon Feder

Informal Systems

Author Note

Ian Thacker https://orcid.org/0000-0002-2492-2929

Correspondence concerning this article should be addressed to Ian Thacker, Department of Educational Psychology, University of Texas at San Antonio, San Antonio, TX 78207. Email: ian.thacker@utsa.edu.

We have registered this study with the University of Texas at San Antonio ethics board (FY20-21-70).

We have no conflicts of interests to disclose.

We would like to acknowledge Rachel Broadway for her contributions in interviewing participants and coding data. We would also like to acknowledge Chelsie Aldridge for her contributions in coding data.

Accepted for publication in the *International Journal of Science Education.*

This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. The final article will be available, upon publication, via its DOI: https://doi.org/10.1080/09500693.2024.2307473

Abstract

Presenting novel numbers about climate change to people after they estimate those numbers can shift their attitudes and scientific conceptions. Prior research suggests that such science learning can be supported by encouraging learners to make use of given benchmark information, however there are several other numerical estimation skills that may also be relevant in this context. This design-based research project aimed to identify specific mathematical skills that might support postsecondary students' learning from novel scientific data. Concurrently, we also developed an open-source online science learning app. In three design iterations, we conducted 22 think-aloud interviews with undergraduate and graduate students at a Hispanic Serving Institution as they estimated climate change data, before being shown the scientifically accepted value. Productive estimation strategies included: tolerance for error, mental computation skills (rounding and arithmetically manipulating given benchmark values), and integration of prior educational and personal experiences. Two cases are presented, the first illustrates a student who used estimation strategies productively and was tolerant of their inaccuracies, the second illustrates a student who reacted negatively to feedback on their inaccuracies. Results implicate principles for integrating mathematics and science learning and showcase a learning intervention that embodies those principles.

Keywords: conceptual change; climate change; design-based research

Despite escalating urgency among scientists around the impacts of climate change due to human activity (Allan et al., 2021), misconceptions about climate change are widespread in the U.S. For example, as of 2021, only 57% of adults in the U.S. agree that "most scientists think that global warming is happening," meaning that the remaining 43% held a serious misconception about the scientific consensus on climate change (Leiserowitz, 2022). Fortunately, several approaches exist to shift climate change misconceptions, and one method that shows promise makes use of pertinent scientific data.

Numbers can be a catalyst for changing minds about science topics. For example, prompting people to estimate just a handful of statistics about climate change and then presenting them with the actual value can shift their attitudes, beliefs, and misconceptions to be more aligned with scientists (Ranney & Clark, 2016). Evidence further suggests that the impact of such an intervention can be enhanced by bolstering targeted numerical estimation skills that support the processing and interpretation of numbers, and that such impacts can be moderated by learner characteristics such as their motivation and learning dispositions (Thacker, 2023; Thacker & Sinatra, 2022). However, it is not clear from this prior research which numerical estimation skills support data interpretation in this context. Further, the interventions created for these prior studies are not widely available to teachers, students, or the general public.

This project aimed to address these gaps by (a) examining specific numerical estimation strategies and student characteristics hypothesized to support conceptual change when university students engage with climate change data, while at the same time, (b) designing an openly accessible online learning app that leverages these skills and characteristics. Specifically, we extend a study conducted by Thacker & Sinatra (2022) by revising their central intervention through several design iterations involving qualitative interviews that highlight estimation strategies that were productive in this context.

Theoretical Framework

Our study aimed to leverage principles of conceptual change and numerical knowledge for the design of a climate change learning intervention. In this section we summarize theories and principles that informed our specific design choices.

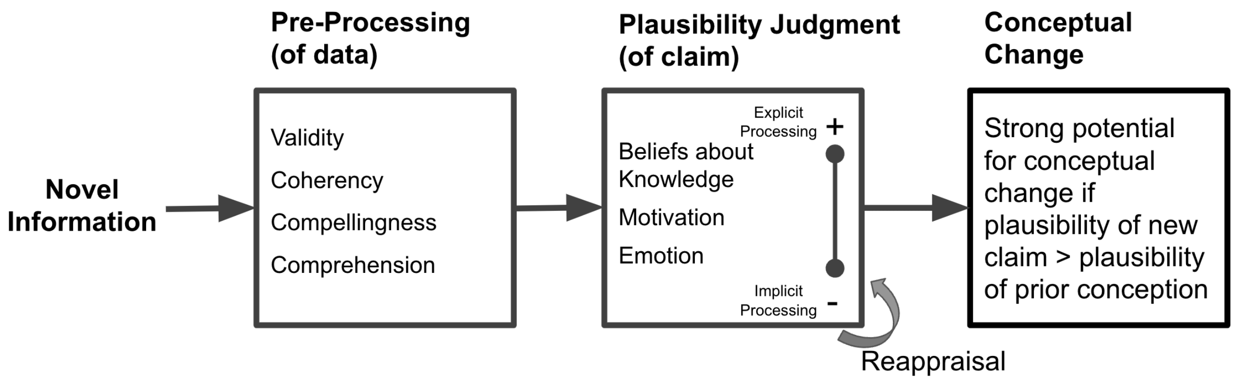
Plausibility Judgments for Conceptual Change

Conceptual change can be defined as a process of restructuring conceptions to be more aligned with the scientific consensus. The Plausibility Judgments for Conceptual Change (PJCC) model posits that novel information can incite conceptual change because it prompts learners to appraise or reappraise the plausibility of their existing beliefs in light of new explanations (Lombardi, et al., 2016). This model predicts that when people encounter novel information, such as novel climate change data, they first process the data for validity (cf., Richter & Maier, 2017) and then judge the plausibility of the conception supported by the new information. The model also predicts that the more explicit and intentional the learner is when processing and considering plausibility, the more likely it is that the learner will change their conceptions.

Based on this model, prompting people to estimate a number before presenting them with the true value can activate prior knowledge and explicit processing. When asked to estimate a number, learners may draw from relevant prior experiences, recall information learned in school or elsewhere, or apply quantitative reasoning skills to arrive at a reasonable estimate (also see the Numerical Driven Inferencing [NDI] framework; Ranney et al., 2001). With this prior knowledge activated, the learner is better prepared to modify their knowledge structures and revise misconceptions when presented with the scientifically accepted numerical value (Dole & Sinatra, 1998; Lombardi et al., 2016; Ranney & Clark, 2016; Siegler, 2016).

Furthermore, when learners are presented with novel information after they estimate it, the extent of their learning and integration of new information depends on a collection of critical factors. The PJCC model integrates ideas from Dole & Sinatra's (1998) model, and predicts that learner characteristics (such as their *comprehension* of the information, perceptions of the *validity*/credibility, and whether the information is *coherent* with their prior understandings) contributes to their initial processing of information, which then predicts their motivation, emotion, and beliefs about learning and knowledge, which subsequently predicts whether they will explicitly consider or reconsider their perceptions of whether claims are plausible, and thus predicts whether conceptual change is likely to occur (See Figure 1 for an illustration of this process model; Dole & Sinatra, 1998; Lombardi et al., 2016). An important implication of the PJCC model is therefore that instructional designs might be most effective when they present novel, comprehensible, and compelling information in a way that might easily be perceived as personally relevant, valid, and credible. Another notable implication of the PJCC is that *personal* perceptions and endorsements (i.e., plausibility judgments) and conceptions of what *scientists* believe (knowledge) are considered to be separate but related constructs. With regards to the topic of climate change, knowledge in this study is measured in terms of whether university students understand that there is a scientific consensus around climate change whereas plausibility is more about the learners' personal endorsements regarding whether climate change seems plausible (e.g., Lombardi et al., 2013; Thacker, 2023; Thacker & Sinatra, 2022). In this study, we used this theoretical model to guide the development of a design to support conceptual change outcomes.

Figure 1.
Conceptual Change Process Model Utilized in This Study.



Note. This conceptual change model was inferred from Dole & Sinatra (1998) and Lombardi et al., (2016).

While there is evidence that presenting people with novel climate change data after they estimate that data can support conceptual change (Ranney & Clark, 2016; Thacker, 2023; Thacker & Sinatra, 2022), little research has intentionally designed interventions with the purpose of seeking out skills to support data comprehension and student motivation. As such, we aimed to create interventions that introduce students to novel data, study specific skills that are important for their comprehension of that information, and explore ways to help contextualize data so learners perceive it as relevant and engaging. To better frame how students process and interpret numerical data in the context of this study, we now turn to literature on numerical estimation and magnitude knowledge.

Magnitude Knowledge

Theory on magnitude knowledge and numerical estimation suggests that specific quantitative reasoning skills may facilitate sense-making processes when people learn from quantities that they encounter. Interpreting numbers and making meaning from them is considered to be a core competency in mathematics and science and involves several relevant quantitative reasoning skills (Booth & Siegler, 2006; Cheuk, 2012; Opfer & Siegler, 2007;

Sasanguie et al., 2012; Siegler & Booth, 2004). Siegler's (2016) Integrated Theory of Numerical Development provides an explanation for how people develop these skills, positing that people develop increasingly accurate understandings of number magnitudes as they connect numbers (e.g., representing escalating global temperatures) to their referents (e.g., global climate change). People make meaning of numbers and build scientific understandings as they connect and compare those numbers to other numbers, ideas, and representations through processes of association and analogy (Siegler, 2016). Numerical estimation is a particularly relevant context in which people tend to apply processes of association and analogy.

Numerical Estimation

One way that learners process and interpret numbers is by estimating whether they seem reasonable (Reys & Reys, 2004). Numerical estimation can be defined as an educated guess of the magnitude of a quantity that can draw from a person's prior experiences and understandings of number and operations (Dowker, 2005). Of the common categories of numerical estimation skills (e.g., computational estimation or numerosity; Reys & Reys, 2004), research on measurement estimation is the most relevant for this study.

Research on *measurement estimation* concerns the explicit estimation of real-world measures (Bright, 1976; Dowker, 2005; Siegler & Booth, 2005; Sowder & Wheeler, 1989) and is useful for understanding factors that help people process real-world quantities and judge whether they are reasonable and valid. Research suggests that people use multiple representations and strategies when estimating measurements. According to a summary of literature (Siegler & Booth, 2005), the strategies that people use during estimation are numerous and range from: applying prior knowledge, rounding digits, using visual representations, making use of given or known information, to flat-out guessing (cf., Joram, 1998). Siegler & Booth, (2005) also posit that people adapt and change strategies based on accuracy feedback and formal instruction, and will select strategies appropriate to the situation if they believe those strategies will lead to more accurate or rapid estimates.

Empirical research suggests that peoples' estimation accuracy and correct judgments of reasonableness improve when they use specific measurement estimation strategies. Three particular strategies have been well explored in the literature: Mentally iterating or scaling given *benchmark values* to better estimate and judge the plausibility of real-world quantities (Brown & Siegler, 2001; Joram et al., 1998; Siegler, 2016), showing a *tolerance for error* and imprecision (Hogan et al., 2004; Shimizu & Ishida, 1994), and *rounding* quantities to make mental computation easier (Alajmi & Reyes, 2010; Reys & Reys, 2004). The use of such strategies is thought to support the integration of novel data and the non-symbolic referents of these numbers (Siegler, 2016) and may therefore support people's comprehension and evaluation of real-world quantities. Below, we discuss each of these three strategies in more detail.

The Benchmark Strategy. The *benchmark estimation strategy* can be defined as the use of a given or known number (called "a benchmark" or "reference point") that can be used to better estimate an unknown number (Reys & Reys, 2004). Prior studies show that presenting undergraduate students with benchmark values can improve their measurement estimation accuracy. For instance, students who estimated the distance between two cities were more accurate when given a handful of benchmark distances between cities as compared with a control (Brown & Siegler, 2001). Similarly, undergraduate students more accurately estimated a wider variety of every-day quantities when provided with benchmarks (Brown & Siegler, 1993, 1996, Friedman & Brown, 2000, Joram et al., 2005), and further, the use of the benchmark strategy by eighth grade students in Kuwait was shown to be a key skill in supporting their correct judgments of reasonableness of given quantities (Alajmi & Reyes, 2010).

However, students do not always spontaneously use benchmark strategies when estimating numbers (Gooya et al., 2011; Hildreth, 1983; Joram et al., 2005) and may need support. Indeed, evidence suggests that instruction on when and how to use benchmark values can increase the frequency of its use, improving estimation accuracy (Joram et al., 2005) and science learning among actively open-minded students who estimate climate change quantities (Thacker, 2023; Thacker & Sinatra, 2022).

Existing interventions have been designed to support the use of benchmarks in measurement estimation. In a design-based research study conducted with four elementary students in the U.S., Hartono & Hartono (2015) investigated how to support the development of students' use of the benchmark strategy for length estimation. The researchers iteratively designed five lessons to shift student strategies from using rulers and precise measurements to finding the lengths of objects using other reference lengths (such as the length of an arm) to assist estimation. The researchers found that students were able to make the shift from precise measurements to using benchmark estimation strategies and identified challenges some of the students encountered during each lesson. Notably, the authors found that explicitly teaching and giving applied opportunities to practice measurement estimation was crucial for student learning.

Tolerance for Error. *Tolerance for error* is a tolerance for imprecision when estimating numbers and valuing "reasonable" answers over "correct" answers and is associated with estimation accuracy and judgments of the reasonableness of numbers. Numerical estimation involves a higher level of imprecision and uncertainty when compared with other mathematical tasks, and maintaining tolerance for that imprecision is important for learners to keep in mind when assessing their own performance on estimation tasks. Individual interviews with university students, adults, and K-12 students suggest that a tolerance for error is a common characteristic among accurate estimators (Reys et al., 1982; Reys & Reys, 2004; Shimizu & Ishida, 1994). However, despite this research showing associations between reasoning and a tolerance for error, no studies have investigated the benefits of teaching a tolerance for error to students before they engage with estimation tasks.

Flexible Rounding. *Flexible rounding* is a set of techniques for rounding digits to ease mental computation, which has been shown to reduce estimation time and improve accuracy (Joram et al., 1998; Reyes & Reyes, 2004; Siegler & Booth, 2005). As with a tolerance for error, while prior research has found correlational evidence that exceptional numerical estimators also tend to use rounding techniques (Reys et al., 1982), there are little to no studies testing whether instruction on the use of rounding strategies can support learning outcomes.

In the current study, we developed an intervention that provides light-touch instruction on underexplored estimation strategies (i.e., the use of benchmark values, tolerance for error, and flexible rounding) to enhance the learning that occurs when university students estimate climate change numbers before being shown the true value.

Current Study

It has been established that numerical estimation is an important mathematical skill that helps students evaluate, interpret, and learn from scientific data. The purpose of this research was to study productive numerical estimation techniques that university students (both graduate and undergraduate) employ when presented with novel climate change data, while concurrently creating learning contexts that emphasize and enhance these strategies.

Therefore, our research questions were:

1. *How can an online intervention be developed to support university students' use of numerical estimation skills and other factors related to processing and interpreting data for the learning of climate change science?*
2. *What numerical estimation skills support the learning that occurs when university students engage with novel climate change statistics?*

Methods

Design-Based Research

To address these questions, we used a design-based research (DBR) approach to guide the development of an online intervention that we call the "Estimation Game" (http://ianthacker.com/design.html). As is characteristic of design-based research, the design, implementation, and revision of the intervention occurred over several iterations as guided by reflection on emerging conjectures about teaching and learning (Anderson & Shattuck, 2012; Bakker, 2019; Cobb et al., 2003). Building from this tradition, Hoadley & Campos (2022) present a process model for design-based researchers who aim to study online learning, and suggest that DBR-researchers should devote attention to (a) grounding initial designs in theory and existing evidence, (b) articulating conjectures, (c) iterating by revising the design and conjectures after testing, and (d) reflecting on the data through the lens of relevant theory, design principles, and learning principles (Hoadley & Campos, 2022). In this way, the iterative process of designing, testing, and redesigning interventions might lead to generalizable conclusions about designed environments, learning theories, and lead to actionable information for practitioners and researchers. As such, we aimed to (a) ground the Estimation Game by rationalizing design elements in terms of the theory of conceptual change and magnitude knowledge (Lombardi et al., 2016; Siegler, 2016) and confirming empirical research (Thacker &

Sinatra, 2022), (b) articulate emerging conjectures about teaching and learning from novel data, (c) iterate on the design and conjectures through piloting the intervention with undergraduate and graduate students, and (d) reflect on the findings to inform theory, design, and generate actionable information.

Furthermore, in addition to creating generalizable information, another outcome of design-based research is the design itself. In the case of this research, that design is an open-access, open-source web app that can easily be shared with educators online.

A Design for Leveraging Estimation Skills

Starting with Thacker & Sinatra (2022) initial design, over three design iterations we revised the intervention to create an online, open-source number Estimation Game with a built-in numerical estimation strategy intervention that can be easily shared with practitioners and the general public online (see Appendix A for a summary of the design and iterations). The intervention prompts the user to estimate climate change numbers before showing them the true value. As noted in the introduction, the initial estimation process is thought to elicit relevant prior knowledge that is then restructured when incorporating the revealed true value (e.g., Lombardi et al., 2016; Ranney et al., 2001; Reys & Reys, 2004). Half of these prompts also included a "hint," (or benchmark value) that could be mathematically manipulated and coordinated with prior knowledge to better estimate the unknown value (Brown & Siegler, 2001; Joram et al., 1998; Siegler, 2016). To improve engagement levels and make the climate change quantities personally relevant (Dole & Sinatra, 1998; Lombardi et al., 2016), participants were also presented with accuracy feedback based on percent accuracy ([estimated value – true value]/true value). The way in which this accuracy feedback was presented evolved over the course of the three design iterations to improve motivation and signal discrepancies between estimates and scientific perspectives (see "Results" section or Appendix A for a description of how accuracy feedback was provided at each stage of the design process). Further, in the second and third design iterations, we incorporated short explanations to contextualize and justify each of the scientifically accepted values that we presented participants with (see Appendix A for a summary of the design trajectory). The rationale for incorporating this additional explanation was to improve coherence with prior knowledge and comprehension of the novel data (Lombardi et al., 2016).

We generated a design in which the researcher was given an option to select whether the Estimation Game would present participants with instruction on numerical estimation strategies prior to estimating. This light-touch instruction consisted of a short text that encouraged participants to draw from their background knowledge and think mathematically when estimating numbers. The initial design emphasized the use of benchmark values by rounding and rescaling them based on one's expectations, followed by a worked example and a check for understanding. Based on data collected from the first iteration, later iterations also presented participants with text to support their tolerance for error by reassuring them that "...it is okay if your estimate is not perfect" as well as support with rounding digits to make mental computation easier (see Supplemental Materials, Appendix SA for full intervention text). Generally speaking, these three estimation strategies (use of benchmark values, tolerance for error, and flexible rounding strategies) were intended to improve comprehension of the true values when presented.

Another factor that guided our design trajectory was our goal to create an experience that is "game-like" by incorporating common game attributes. Namely, of nine common attributes identified by game developers and gamers (i.e., action language, assessment, challenge, control, environment, interactivity, immersion, and goals/rules; Bedwell et al., 2012), the estimation intervention we developed in this study align with many (though not all) of them. Though the intervention is very short in duration and *does not* involve prominent elements of fantasy/story-based elements, human interaction, immersion, control, or adaptive levels of challenge, the game *does* include a clear action language (entering numbers), goals (accuracy in estimates of unknown quantities), assessment (accuracy scores), and environmental context (global climate change). As such, we refer to the central intervention as the "Estimation Game" despite certain game attributes that could use strengthening in future iterations.

Data Sources

Participants and Procedure

The lead author and a graduate student research assistant conducted 22 audio and video recorded "think-aloud" interviews (Desimone & Le Floch, 2004) in a one-on-one setting via Zoom. Student participants attended a large Hispanic serving research university in the Southern USA and identified as undergraduate students (36%), graduate students (64%), Female (77%), Hispanic/Latino (50%), White (27%), Black (5%), American Indian (5%), two or more races (14%). Ten students participated in the first iteration, nine in the second, and three in the third. All interviews were proctored by either the first author of this study or a graduate research assistant. The first author identified as White and male and was faculty at the same institution as the students in the current study. The graduate student interviewer identified as White, female, and was a graduate student attending the same institution.

After students completed an informed consent questionnaire, the interviewer acquainted students with the interview procedure, and asked participants to "think aloud" while completing a pretest, engaging with the Estimation Game, and a posttest (see Supplementary Materials, Appendix SB for all survey materials and Appendix SC for the interview protocol). To acquaint students with the interview process and to build rapport, the interviewer modeled the interview process by “thinking aloud” while responding to an example survey item about TV watching habits; participants then followed suit by demonstrating their understanding of the procedure by thinking aloud while responding to the same item. They then were asked to continue to think aloud while interacting with the Estimation Game and responding to items in a pretest and posttest survey. The interviewer would occasionally remind students to think aloud if there was a long pause and would occasionally ask followup questions if they had issues interpreting survey items or wording of the Estimation Game text (see Appendix SC for the full interview protocol). After engaging with the Estimation Game that required estimating 12 quantities, participants were prompted to answer five questions about the nature of the experience (e.g., “What would you change about the look and feel of the game?”) before completing the posttest questionnaire.

Survey Materials

The pretest and posttest survey assessed students’ climate change knowledge (see Supplementary Materials, Appendix SB for all survey materials) as well as additional unfinished scales that were in the process of being validated for a separate study (Thacker, 2023). The climate change knowledge measure consisted of seven items adapted from Lombardi et al., (2013) intended to capture students’ understanding about the scientific consensus around climate change. These items prompted participants to indicate the extent to which they believe that *scientists* agree with certain statements on a five-point agreement scale (e.g., whether scientists agree that “Greenhouse gas levels are increasing in the atmosphere”). This knowledge measure was selected because it was found to have strong content validity and reliability when used with undergraduate students (Lombardi et al., 2013). In the current sample, the pretest and posttest had a Cronbach’s alpha of .52 and .68 respectively. We also calculated the difference in mean scores (mean posttest – mean pretest), and used the median of this gain score to split the sample into two balanced groups of 11 students (which we refer to hereafter as "high gain" and "low gain" groups). Participants also completed a demographics questionnaire at posttest. Descriptive statistics for the knowledge survey are presented in the Supplementary Materials (Table S1).

Analysis

The analysis occurred in four waves: three waves after each of the three design iterations and the fourth after all data was collected. For each design iteration, recordings from interviews were transcribed and analyzed. Interviewers wrote analytical memos based on an open analysis of each interview transcript. Conclusions that were derived from analyses were then used to modify and improve the Estimation Game and to test emerging conjectures (Bakker, 2019; Cobb et al., 2003; Hoadley & Campos, 2022; Design-Based Research Collective, 2003) to address our two research questions. These research questions focused on supporting students in applying numerical estimation skills to help them learn from climate change quantities, and identifying skills and student characteristics that were relevant for this integration of mathematics and science knowledge. Throughout this process, our team met weekly to discuss the analyses, insights, and conclusions, and to plan ways to modify the Estimation Game design before the next iteration.

After three design iterations, all transcribed recordings were collectively revisited and open-coded by the lead author and a graduate research assistant for varying dimensions of student thinking (Corbin & Strauss, 2004; Saldaña, 2010). A codebook was collaboratively developed around themes that emerged. All themes centered around how students made sense of climate change numbers, with major categories being (a) strategies employed when estimating climate change quantities (wild guess, mental computation, prior knowledge), and (b) emotions and attitudes expressed when reacting to accuracy feedback (tolerance for error and negative/positive reactions to accuracy feedback).

Two graduate student coders then independently coded all data over the course of four weeks and met for weekly meetings to calibrate and revise the codebook definitions (Saldaña, 2013; see Appendix SD for full codebook). We retained all codes with interrater agreement greater than 95% at the paragraph level—which is thought to be an appropriate indicator of interrater reliability for low-incidence codes (McHugh, 2012); though we also report Cohen's kappa coefficients (κ) which account for agreement due to chance. Interrater reliability was sufficient for the following codes: tolerance for error (96.6% agreement, $\kappa = .81$), negative reaction to accuracy feedback (95.2% agreement, $\kappa = .65$), positive reaction to accuracy feedback (99.7% agreement, $\kappa = .82$), and wild guess (95.2% agreement, $\kappa = .64$). However, sub-codes that fell under the categories of "prior knowledge" and "mental computation" were not reliable at the 95% level. As such, all "prior knowledge" subcodes were collapsed into a single parent code which increased its interrater agreement to 98%, $\kappa = .86$. We also revised the codebook to refine and clarify all mental computation subcodes. A third coder (second author of this manuscript) then used this revised codebook to recode all transcripts, paying specific attention to identifying mental computation strategies. During weekly meetings, our team discussed each instance of the mental computation sub-codes, resolved any disagreements, and came to a consensus on the refined codes (i.e., interrater agreement was 100% after resolving differences for all codes).

To identify estimation strategies that may have supported learning gains, we divided the participants into those with above- and below-median knowledge gain scores on the knowledge assessment and compared the frequency of their use of estimation strategies thus splitting the sample into two balanced groups of $n = 11$.

Results

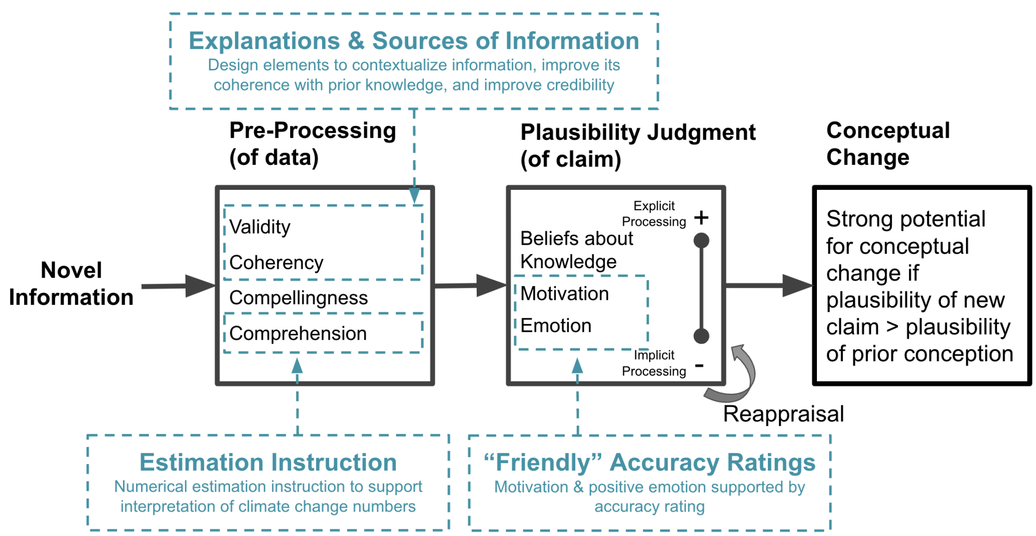
Survey results revealed that many participants initially held misconceptions at pretest that improved at posttest. For example, 32% of students initially disagreed with the proposition that "scientists believe that average sea level is increasing," suggesting that these students held a misconception about the scientific consensus around climate change. Notably, this rate is near the national average of 43% of adults in the United States who also disagree with the statement (Leiserowitz, 2022). Pretest climate change knowledge scores improved from 73% correct at pretest to 94% correct at posttest (also see Supplementary Materials Table S1).

RQ1: Design of an Open-Access Estimation Game

In this section, we report on the trajectory of the three design iterations, our conjectures about learning that informed those modifications, and the rationale behind each modification (for a summary of the design trajectory, see Appendix A; for alignment with our theoretical model, see Figure 2).

Figure 2.

Conceptual Change Process Model Flagged with Relevant Design Elements



The first design iteration redeployed two core design features of the original intervention created by Thacker & Sinatra (2022). First we asked participants to estimate novel climate change numbers, and then introduced them to the actual values. As noted in the introduction, this feature was motivated by the theory positing that estimating a number can elicit prior knowledge (Ranney et al., 2001) which can be restructured after novel or incongruous evidence is presented (Dole & Sinatra, 1998; Lombardi et al., 2016). Second we included instructional support around use of the benchmark strategy as a way to help students engage more deeply with climate change quantities and to support their comprehension and learning of what those numbers refer to (Siegler, 2016). See iteration 0 in Appendix A, "Summary of Design Iterations" for illustrations of these core design features.

Four improvements were introduced on top of the core features of the original design for the initial iteration of this study. First, the intervention was implemented as an open-source, and open-access web app. Second, we gave learners the option to view the sources of the information in order to build a sense of source credibility, which was expected to lead to more explicit plausibility reappraisals and increase the likelihood of conceptual change (Lombardi et al., 2016). Third, we generated personalized accuracy feedback that was presented to learners after each estimate to provide useful feedback and improve engagement. This was also expected to lead to explicit plausibility reappraisals and catalyze conceptual change (Lombardi et al., 2016). Accuracy feedback was determined based on percent accuracy calculation ([Estimated value - True value]/True value), and one of four accuracy thresholds (<10% = Very accurate, <20% = Accurate, <30% = Inaccurate, >30% = Very inaccurate) were displayed in a box that had a color ranging from green (accurate) to red (inaccurate), depending on the level of accuracy. Fourth, after estimating all 12 climate change numbers, the accuracy ratings, as well as all estimates, true values, and sources of information for each question were presented on a summary page to act as a final "scoresheet" for participants, encouraging them to reflect on the experience. See iteration 1 in Appendix A, "Summary of Design Iterations" for illustrations of these improvements.

Analysis of the first round of interviews using the initial design revealed comments from students that then informed subsequent design revisions. A critical signal from these comments was given at the conclusion of the study: When asked "What would you change about the look and feel of the game?", several students reported perceiving the accuracy feedback display as "aggressive." For example, when introduced to the summary page displaying all student estimates and true values highlighted, a student remarked "That's a little overwhelming… It's a little aggressive, I'm not going to lie to you." Indeed, our analysis of the interviews generally indicated that the jarring shade of red indicating negative performance feedback appeared to stand out to students more than the focal climate change numbers. Such negative performance feedback sometimes appeared to frame students' perceptions of the scientific values in terms of whether they "got the answer right," rather than encouraging them to focus on the meaning of those scientific values. For example, a student remarked on their accuracy by saying, "Oh, I was way off" and another "I'm getting these all wrong" without commenting on *why* they were inaccurate. Further, participants also reported that the true value, when presented, was not easy to comprehend without context, and recommended that we include an explanation to help contextualize the information. For example, upon concluding the game and asked what they would change about the experience, a student recommended that we might consider providing "…a blurb with information from like what the accurate information would have come from… Like maybe you could pull something from that source, like where they got that information from." Such feedback from students was crucial for subsequent design revisions.

For the second iteration, we redesigned the Estimation Game based on the recommendations and findings from the first iteration by (a) reducing the salience of negative feedback, (b) including additional instruction around tolerance for error, (c) adding evidence and explanations to better contextualize climate change data, and (d) improving the "look and feel" of the game to be more game-like. Regarding the first change (a), we reserved all performance feedback until the end of the experience, where we displayed it on the summary page. This aimed to reduce the salience of the negative performance feedback, which seemed to be negatively impacting students' motivation as they progressed through the game (see Table 1). The second change, (b), added text to the estimation instruction to emphasize and improve students' tolerance for error, aiming to reduce anxiety around "getting the answer wrong," and refocus attention on the novel evidence. The third change, (c), added explanations and additional evidence to help contextualize each true value in order to increase the comprehensibility and credibility of the information. The fourth change, (d), revised the "look and feel" of the game to be more game-like by drawing from a retro 8-bit aesthetic template (Rikko, n.d.).

Findings from the second iteration of interviews revealed that participants appreciated the context explanations. For example, at the conclusion of an interview, one student commented that the explanations were helpful, "They explained it. Like, what the answer was, and what's causing it, and how [we] can fix it going forward." Students also expressed curiosity about the accuracy of their estimates and shared that they preferred to have more immediate feedback. For example, when asked about what they would change about the game, one student mentioned "This results screen is fine, but it's not like I'm getting points for it, I'm just seeing how accurate or inaccurate I was. And it doesn't reward you." This comment suggests that, because results were presented only at the conclusion of the game, students didn't view them as central to the experience. We also noticed that students might benefit from a cue to potentially reappraise the assumptions informing their estimates. As a compromise between the first iteration ("aggressive" feedback) and the second iteration (no immediate feedback), we chose to reintroduce the accuracy feedback feature of the game, but in a way that emphasized positive aspects of their performance rather than negative; we incorporated feedback that was more "friendly."

The third iteration mostly included revisions to how accuracy feedback was presented. Stars were introduced to represent accuracy, with the same accuracy thresholds as presented in the first iteration (four stars for an estimate within 10% of the true value, three for within 20%, two for within 30%, one for outside of 30%, and no stars for indicating an incorrect direction). Interviews from this iteration revealed that participants generally reacted positively to the "friendly" accuracy feedback. One student, when asked what they would change about the look and feel of the app, said "I liked the stars. They weren't too flashy… I don't think you need to change anything about it." Though, during interviews, we noticed that some students shared that they believed they received the minimum score when their accuracy rating was a single star. For example, when reflecting on the game while viewing the final results page, the interviewer noticed that a student seemed discouraged by their results, despite earning a non-zero score for every question, and thus asked the student "What do you think is the minimum number of stars you can get?" to which she responded "The minimum number, I guess, is one," suggesting that the student was under the impression that they received a minimum score for many items. As such we made a final revision to the design by adding a fifth star threshold, with one star being the minimum.

RQ2: Estimation Strategies that Support Learning

To address our second research question, focused on identifying estimation skills among university students that support their science learning, we compared strategies used by participants with above- and below-median learning gains. Frequency counts for the number of participants that used different estimation strategies, the total number of times that they used the strategies, and strategy use grouped by knowledge gain are presented in Table 1. Definitions and excerpts illustrating each code can be found in the Supplemental Materials (Appendix SD).

All students drew from their *prior knowledge* such as referencing educational, personal, or social media experiences (e.g., "[based on] the documentaries I've seen in my life, they been saying things like [temperature] generally increases"). All students used several *mental computation strategies* when estimating quantities, which included *flexible rounding* of digits (e.g., "I [rounded to] 194 because [halving] 195 will result in a point five calculation, and there's not really half a country, so I just rounded down because, you know, down is less") and three forms of *benchmark estimation*. Regarding the use of given benchmarks, participants either used benchmarks vaguely (e.g., "I used the hint"), used arithmetic (e.g., by using repeated addition ["So we'll add another third to that so 1.24 plus point three three I'll say is 1.57"] or proportional reasoning ["I'll put 1.5% because it's half of 3%"]), or extrapolated climate impact trends by using given information to project into the future ("so if we take into consideration that increase [given in the hint] and we use the same logic, so 0.7 per every 50 years.").

Table 1.
Estimation Strategies Employed by Graduate and Undergraduate Students While Estimating Climate Change Numbers

	Total Number of *participants* who used strategy at least once	**Total** Number *of times* participants used this strategy	**Low-Gain** (% of time that this strategy was used among low-gain participants)	**High-Gain** (% of time that this strategy was used among high-gain participants)	**p-value** From one-sample, two-tailed, proportion z-test

Mental Computation	22	119	45%	55%	.230
Benchmark Estimation (unspecified)	11	24	55%	45%	.624
Benchmark Estimation (Explicit use of arithmetic)	17	53	40%	60%	.145
Benchmark Estimation (Extrapolation)	14	27	41%	59%	.349
Flexible Rounding	12	15	53%	47%	.816
Prior Knowledge	22	179	46%	54%	.284
Tolerance for Error	**18**	**53**	**38%**	**62%**	**.081~**
Negative Reaction to Accuracy Feedback	12	26	65%	35%	.126
Positive Reaction to Accuracy Feedback	**16**	**84**	**36%**	**64%**	**.010***
Wild Guess	17	74	58%	42%	.167

Note. "Total" column indicates the number of *participants* who were identified at each code at least one time (i.e., number of participants who exhibited each code out of 22). "High and Low-Gain %" represents row percentages of coding references by knowledge gain groups, (i.e., number of [high, low] gain coding references ÷ total references). The Prior Knowledge code is a parent collapsed across all related subcodes (knowledge from news, social media, school-based, personal experiences, or vague use of prior knowledge) to improve inter-rater agreement. P-value calculation based on two-tailed, one sample test of proportions as to test whether the proportion of times a high-gain student used a strategy was different than 0.5.
$\sim p < .1$; $*p < .05$

We also found that participants with above-median learning gains tended to show greater *tolerance for error* (e.g., "Oh, my gosh that's not bad. 1.6 in comparison to five it's bad, but like in comparison [to my estimate] that's not bad at all. I'm okay with that."), at marginally significant levels ($p = .081$). This is consistent with prior research finding that greater tolerance for error and willingness to make calculations that may not be accurate is characteristic of good numerical estimators (cf., Reys et al., 1982; Shimizu & Ishida, 1994). Given this tolerance for imprecision, it is not surprising that this group also shared significantly more *positive reactions*

(p = .010; e.g., “Yeah I got that pretty much right. Overwhelming majority, 9 out of 10”) to accuracy feedback when their estimates were not accurate.

Two Contrasting Illustrative Vignettes

To exemplify how specific estimation strategies may have supported student learning, and how accuracy feedback elicited various responses, we present excerpts from two interviews, both of whom coincidentally received exactly the same knowledge score at pretest. The first was with an undergraduate student in a teacher preparation program who was identified as “high-gain” because she had above-median improvements on the climate change knowledge measure, having improved from 2.86 of 5 at pretest to 3.72 of 5 at posttest. This student self-identified as female, 22 years old, multiracial and was selected for presentation because she represents the case of a student who effectively applied quantitative reasoning skills to estimate and interpret climate change numbers presented in the Estimation Game. This student was interviewed by the lead author during the first iteration of interviews. The interviewer had not met the student prior to the study.

After introductions, signing informed consent, and completing the pretest survey, the student was thinking aloud when she read the instructions to the item, “Of 195 countries in the world how many are committed to climate action?” She then noted that she was “gonna round down to 194.” When prompted by the researcher to explain why, she said,

> I [rounded to] 194 because [halving] 195 will result in a point five calculation, and there’s not really half a country, so I just rounded down because, you know, down is less... Half of 194 is 97, but I’m going to put 42 countries because it’s less than half of 195. [*She then enters 42 and clicks to show the scientifically accepted answer, revealing that 175 of 195 countries are committed to climate action.*] No! I mean, yes! But no, I mean yes! So

> that's more than half. That's significantly more than half. Wow, that genuinely surprises me a lot. I did not know that. I really thought that a lot of the countries were not committed to climate action. This is a good statistic. I'm happy with this. I mean, I'm sad that I'm wrong, but I'm happy that I'm wrong at the same time.

This excerpt illustrates the flexible approach to working with imperfect calculations that were characteristic of high-gain participants. Notably, this student drew from her prior knowledge and expectation that the world is not very supportive of climate action and performed a few casual "back of the envelope" computations using the given number (i.e., rounding to an even number to ease halving, and then implicitly rounding and halving again). These casual arithmetic manipulations seemed to support students in making meaning of the numbers. Then, when shown the true value, this student noted that, though she was not particularly happy to learn that her estimate was inaccurate, the meaning of the scientifically accepted value was most salient. Notably, she referred to the quantity in terms of her calculations noting that it was "significantly more than half."

In contrast, we found that students identified as low-gain learners were generally less willing to manipulate given numbers and attended more to the accuracy of their estimate when compared with the meaning associated with the true value. The following excerpt comes from an undergraduate student that was identified as having below-median learning gains (2.86 of 5 at pretest, 3.29 of 5 at posttest), who seemed to be distracted by the accuracy feedback. This student initially believed that climate change is real and self-identified as female, Hispanic/Latino, and 20 years old.

The student was thinking aloud when responding to an estimation problem, "How many square kilometers of Arctic ice cover was there in September of 2017? Hint, in September of 2010 the ice cover was 6.54 million square kilometers." In response, she said,

> Okay, so it's asking me how many square kilometers of ice was covered in September of 2017. There were 6.5 covered in 2010 so that's seven years that has passed so then… I would totally guess here because I don't know anything about this, I don't even know like that's normal or not normal… so I would say something like 8 million, which is probably so wrong. [*She then enters 8 million* km^2*, revealing the scientifically accepted answer of 4 million* km^2] Four! Okay, so I was supposed to like half that. I tried. I guessed on that.

This excerpt illustrates the case of a student who seemed to experience anxiety over inaccuracy feedback, and essentially guessed without drawing from prior knowledge or considering the meaning of the "true value" when presented with it. Note that, in this excerpt, this student estimated that ice cover actually *increased* over time, despite previously indicating that she agreed that "most of the world's glaciers are decreasing in size" on the knowledge pretest. This seems to show that the student was distracted by the accuracy feedback to the point of losing sight of the central information being communicated in these items, one of the reasons why we revised the accuracy feedback to be more "friendly" in later iterations. Another explanation for this student's response is that she may have experienced some level of math anxiety, and chose to rush through the game as to avoid engaging deeply in quantitative reasoning (for additional evidence on math anxiety and its negative relationship to learning from this intervention, see Thacker, 2023).

Discussion

We explored skills that support the learning through an intervention involving numerical estimation of climate change data. Over the course of three design iterations, we developed an open-source online technology intended to support numerical estimation skills, processing, and interpretation of novel data for the learning of climate change science. Along with lessons learned from the design process, we also identified estimation strategies used by university students as they estimated climate change numbers and made mathematical and personal meaning from the true values. Following, we discuss our findings and provide implications for research and instruction.

An Estimation Game Introducing Novel Data and The Role of Accuracy Feedback

We intended to design technology that would present university students with novel climate change data after asking them to estimate a quantity. Particular attention was given to integrating design features used in earlier research (e.g., Thacker & Sinatra, 2022) to elicit their prior knowledge and engage students in explicit consideration of novel data. Three design iterations revealed the need to modify the feedback system in order to frame accuracy feedback in such a way that it deemphasized performance outcomes and emphasized the importance of integrating the meaning behind novel data. Namely, we modified the feedback system to present "friendly" gold stars that emphasized estimation accuracies rather than "aggressive" red bars that emphasized inaccuracies. This change, alongside explanations to contextualize novel data for improved coherence with prior knowledge, text to promote tolerance for error, and sources to promote information credibility, appeared to create an environment in which students felt more comfortable estimating and making sense of climate change numbers.

These findings highlight important tradeoffs to be considered when providing accuracy feedback during estimation tasks. On the one hand, accuracy feedback seemed to serve as a cue to direct students' attention towards discrepancies between their prior understanding of climate change quantities and the true values, as well as to help personalize information and create a more engaging and game-like environment for students. On the other hand, such feedback needed to be carefully designed to avoid over-emphasizing negative aspects of performance, given that such feedback had the potential to elicit negative responses from students and lead them to devote attention to the fact *that* they were inaccurate, rather than attending to *why* they were inaccurate. Future research might incorporate experimental designs that test the impacts of different forms of estimation accuracy feedback on various learning and motivational outcomes.

Our findings also suggest that dually supporting comprehension of novel data as well as attending to student motivation is important for learning. Our design trajectory was guided by student responses to design decisions that were informed by the Plausibility Judgments for Conceptual Change model (Lombardi et al., 2016; also see Figure 2). Namely, design iterations were directed toward supporting students' processing of information by leveraging mathematical skills that enhance the *comprehensibility* of novel data, supporting *coherence* with prior knowledge, and emphasizing *credibility* of the information. We also included elements that support plausibility perceptions of climate change, by facilitating motivation, engagement, and thus driving explicit reflection on the core claims underlying the climate change data (i.e., that there is a scientific consensus). Future research might investigate the interrelationships between data processing, plausibility judgments, and conceptual change proposed by this model more systematically. For example, we did not measure learners' personal perceptions of plausibility, which might be attended to and measured in future studies.

Tolerance for Error: A Key Skill Among Above-Median Learners

We found that students with above-median climate change learning gains were more tolerant of error and had more positive reactions to accuracy feedback compared to students with below-median gains. This finding is consistent with prior research suggesting that tolerance for error is an important disposition that facilitates learning during numerical estimation activities and that not all students are prepared with this tolerance (Reys et al., 1982; Reys & Reys, 2004; Shimizu & Ishida, 1994). High-gain participants also tended to react more positively to accuracy feedback, regardless of whether they were accurate or not. Future research might investigate whether it is possible to promote and improve tolerance for error on estimation tasks.

Limitations

This study has necessary limitations. First, we prioritized qualitative, design-based research methods to inform the development of online learning technology, and thus had a sample size that was underpowered for testing differences between high- versus low-gain groups. Future research might make use of larger sample sizes to improve the power of such analyses. Second, the students who participated in this study were graduate and undergraduate students from a Hispanic Serving Institution in the U.S.A., and all reported that they thought it was plausible that "climate change is happening." As such, the participants may not be representative of the national average of adults with regard to climate change beliefs or demographic characteristics. Third, it should be noted that this study represents cross-sectional data and that relationships between high- and low-gain groups are correlational. Future studies might make use of experimental methods to enable causal inference. Fourth, although the main goal of this study was to support and measure knowledge revision, we also developed the Estimation Game with the intention of targeting factors related to students' personal perceptions of plausibility.

However, we did not consistently measure students' personal endorsements about climate change, which might be addressed in future research.

Implications for Teaching

As is characteristic of design-based research studies, one of the main outcomes of this study was a design that could be easily shared with students and teachers. As such, the web app created for this study (http://ianthacker.com/design.html) might be useful for teachers of science who wish to integrate mathematical thinking into their lessons, or teachers of mathematics who wish to integrate relevant scientific information into their instruction. The technology could also be easily adapted to promote group activities that engage students in important conversations about climate change that are informed by relevant data and evidence.

Findings from this study also suggest that instructors who wish to support science learning through casual estimation of real-world quantities might assure students that, when estimating, it is okay to tolerate some error in computations. They may also provide feedback that de-emphasizes performance outcomes and highlights meaning. Future research might investigate more closely the effects of supporting such mathematical reasoning skills on interpreting scientific meaning.

We also found that some students who seemed averse to quantitative reasoning spent less attention to interpreting scientifically accepted values when presented with them and had lower learning gains. Future designs might modify the intervention to consider individuals with higher math anxiety groups by deemphasizing the mathematics instruction, and better emphasizing non-quantitative aspects of the game such as connecting to relevant background knowledge or including more game-like elements such as a deeper and more engaging storyline.

Conclusions

Despite the intensifying effects of climate change, misconceptions about climate change continue to persist in the U.S. The design we created for this study aimed to alleviate such misconceptions among university students by integrating several ideas from the research on numerical estimation and conceptual change. By engaging students in this unique context—in which they apply quantitative reasoning skills to make meaning of numbers, tolerating their own inaccuracies, and reevaluating claims about the world—students may learn not only about climate change, but about the integrated nature of math and science.

Ethics Statement

We have registered this study with the University of Texas at San Antonio Institutional Review Board (FY20-21-70).

References

Alajmi, A., & Reys, R. (2010). Examining eighth grade students' recognition and interpretation of reasonable answers. *International Journal of Science and Mathematics Education*, 8(1), 117–139.

Allan, R. P., Hawkins, E., Bellouin, N. & Collins, B. (2021). *IPCC, 2021: Summary for Policymakers.* In V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., et al., (Eds.) Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press: Cambridge, UK.

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research?. *Educational researcher*, *41*(1), 16-25.

Bakker, A. (2019). *Design Research in Education: A Practical Guide for Early Career Researchers.* Routledge: New York, NY.

Bedwell, W. L., Pavlas, D., Heyne, K., Lazzara, E. H., & Salas, E. (2012). Toward a taxonomy linking game attributes to learning: An empirical study. *Simulation & Gaming, 43*(6), 729-760.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189

Bright, G. W. (1976). Estimation as part of learning to measure. In D. Nelson & R. E. Reys (Eds.), *Measurement in School Mathematics: 1976 Yearbook* (pp. 87–104). Reston, VA: NCTM.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: a framework for understanding real-world quantitative estimation. *Psychological Review, 100*(3), 511.

Brown, N. R., & Siegler, R. S. (1996). Long-term benefits of seeding the knowledge base. *Psychonomic Bulletin & Review, 3*(3), 385-388.

Brown, N. R., & Siegler, R. (2001). Seeds aren't anchors. *Memory & Cognition*, *29*(3), 405–412. Cheuk, T. (2012). *Relationships and convergences found in the Common Core State Standards in mathematics (practices), Common Core State Standards in ELA/literacy (student portraits), and a framework for K-12 science education (science & engineering practices)*. http://nstahosted.org/pdfs/ngss/explanationofvenndiagram.pdf

Carey, S. (2000). The origin of concepts. *Journal of Cognition and Development, 1*(1), 37-41.

Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13.

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Sage publications.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5-8.

Desimone, L M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*(1), 1-22.

Dole, J. A., & Sinatra, G. M. (1998). Reconceptalizing change in the cognitive construction of knowledge. *Educational Psychologist, 33*(2-3), 109-128.

Dowker, A. (2005). A good guess: Estimation and individual differences: Implications for psychology, neuroscience and education. In *Individual differences in arithmetic* (pp. 132–151). Psychology Press.

Friedman, A., & Brown, N. R. (2000). Updating geographical knowledge: Principles of coherence and inertia. J*ournal of Experimental Psychology: Learning, Memory, and Cognition, 26*(4), 900–914. https://doi.org/10.1037/0278-7393.26.4.900

Gooya, Z., Khosroshahi, L. G., & Teppo, A. R. (2011). Iranian students' measurement estimation performance involving linear and area attributes of real-world objects. *ZDM, 43*(5), 709-722.

Hildreth, D. J. (1983). The use of strategies in estimating measurements. *The Arithmetic Teacher, 30*(5), 50–54. https://doi.org/10.5951/AT.30.5.0050

Hoadley, C., & Campos, F. C. (2022). Design-based research: What it is and why it matters to studying online learning. *Educational Psychologist, 57*(3), 207-220.

Hogan, T. P., Wyckoff, L. A., Krebs, P., Jones, W., & Fitzgerald, M. P. (2004). Tolerance for error and computational estimation ability. *Psychological Reports, 94*(3_suppl), 1393-1403.

Hartono, R., Ilma, R., & Hartono, Y. (2015). Supporting the development of students' reference points for length estimation. Retrieved 11.28.2022 from https://repository.unsri.ac.id/6846/. Joram, E., Gabriele, A. J., Bertheau, M., Gelman, R., & Subrahmanyam, K. (2005). Children's use of the reference point strategy for measurement estimation. *Journal for Research in Mathematics Education, 36*(1), 4-23.

Joram, E., Subrahmanyam, K., & Gelman, R. (1998). Measurement estimation: Learning to map the route from number to quantity and back. *Review of Educational Research*, *68*(4), 413–449.

Leiserowitz, A., Maibach, E., Rosenthal, S., Kotcher, J., Carman, J., Neyens, L., Myers, T., Goldberg, M., Campbell, E., Lacroix, K., & Marlon, J. (2022). *Climate Change in the*

American Mind, April 2022. Yale University and George Mason University. New Haven, CT: Yale Program on Climate Change Communication.

Lombardi, D., Nussbaum, E. M., & Sinatra, G. M. (2016). Plausibility judgments in conceptual change and epistemic cognition. *Educational Psychologist*, *51*(1), 35–56.

Lombardi, D., & Sinatra, G. M. (2012). College students' perceptions about the plausibility of human-induced climate change. *Research in Science Education*, *42*(2), 201–217.

Lombardi, D., Sinatra, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction*, *27*, 50–62.

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive psychology, 55*(3), 169-195.

Ranney, M., & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science*, *8*(1), 49–75.

Ranney, M. A., Shonman, M., Fricke, K., Lamprey, L. N., & Kumar, P. (2019). Information that boosts normative global warming acceptance without polarization: Toward JS Mill's political ethology of national character. In D. Wilkenfeld & R. Samuels (Eds.) *Anthology on the experimental philosophy of science* (pp. 61-96). (In Bloomsbury's Advances in Experimental Philosophy series). New York: Bloomsbury

Ranney, M., Cheng, F., Nelson, J., and Garcia de Osuna, J. (2001). *Numerically driven inferencing: A new paradigm for examining judgments, decisions, and policies involving base rates.* Paper presented at the Annual Meeting of the Society for Judgment & Decision Making.

Reys, R. E., & Reys, B. J. (2004). Estimation in the mathematics curriculum: A progress report. *Beyond Written Computation*, *29*, 101–112.

Reys, R. E., Rybold, J. F., & Bestgen, B. J. (1982). Processes used by good computational estimators. *Journal for Research in Mathematics Education*, *13*(3), 183–201.

Richter, T., & Maier, J. (2017). Comprehension of multiple documents with conflicting information: A two-step model of validation. *Educational Psychologist, 52*(3), 148-166.

Rikko, B.C. (n.d.). NES.css: NES-style CSS framework. GitHub Repository, Retrieved on 12/13/2022 from https://nostalgic-css.github.io/NES.css/

Rinne, L. F., Ranney, M. a, & Lurie, N. H. (2006). *Estimation as a catalyst for numeracy: Micro-interventions that increase the use of numerical information in decision-making.* Proceedings of the Seventh International Conference of the Learning Sciences, 571–577.

Saldaña, J. (2021). *The coding manual for qualitative researchers.* Sage Publications.

Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2012). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology, 30*(2), 344-357.

Shimizu, K., & Ishida, J. (1994). The cognitive processes and use of strategies of good Japanese estimators. In R. E. Reys & N. Nohda (Eds.), Computational Alternatives for the twenty-first century: Cross-cultural perspectives from Japan and the United States (pp. 161-178). Reston, VA: National Council of Teachers of Mathematics.

Siegler, R. S. (2016). Magnitude knowledge: The common core of numerical development. *Developmental Science*, *19*(3), 341-361.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development, 75*(2), 428–444. https://doi.org/10.1111/j.1467-8624.2004.00684.x

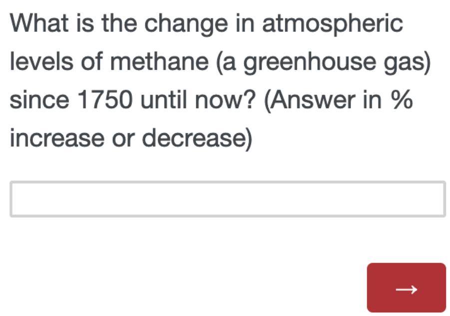
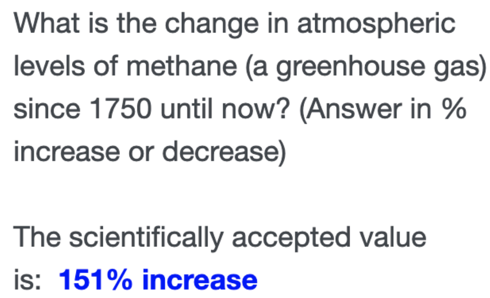
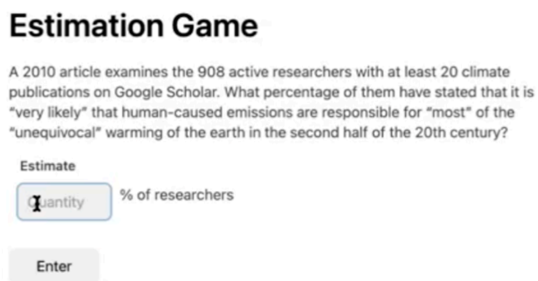
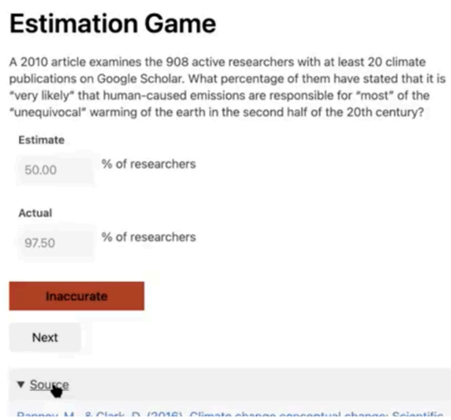
Siegler, R. S., & Booth, J. L. (2005). Development of numerical estimation: A review. In J. I. D. Campbell, (Ed.). (2005), *Handbook of mathematical cognition*. New York: Psychology Press.

Sowder, J. T., & Wheeler, M. M. (1989). The development of concepts and strategies used in computational estimation. *Journal for Research in Mathematics Education*, 130–146.

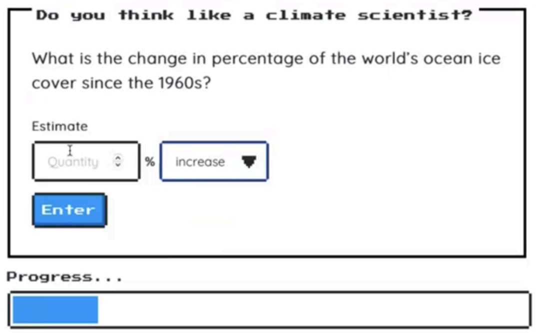
Thacker, I. & Sinatra, G. M. (2022). Supporting climate change understanding with novel data, estimation instruction, and epistemic prompts. *Journal of Educational Psychology, 114*(5), 910–927.

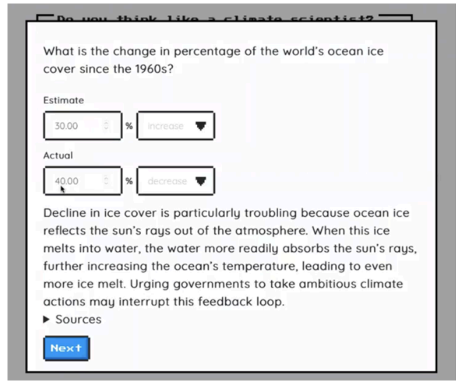
Thacker, I. (2023). Climate change by the numbers: Leveraging mathematical skills for science learning online. *Learning & Instruction, 86*, 101782. https://doi.org/10.1016/j.learninstruc.2023.101782

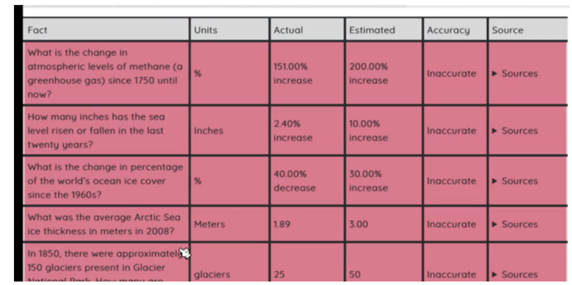
Appendix A. Summary of Design Iterations

<table>
<tr><th>Iteration</th><th>Prompt to Estimate Climate Change Number</th><th>Presentation of True Value</th><th>Final Page
(Summary of All Estimates and True Values)</th></tr>
<tr><td rowspan="2">0</td><td>
</td><td>
</td><td>N/A</td></tr>
<tr><td colspan="3">Thacker & Sinatra (2022) original version of the Estimation Game, implemented via Qualtrics with no accuracy feedback. Initial results from this study revealed improvement in knowledge but no change in plausibility judgments, leading subsequent design revisions to focus on improving climate change plausibility.</td></tr>
<tr><td rowspan="2">1</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">For the first redesign iteration, three key design features were added. This included (a) an option to view the source of the information to improve source credibility and plausibility of claims, (b) five levels of accuracy feedback (with "Very Accurate" for within 10% of the true value, "Accurate" for within 20%, "Less Accurate" for within 30%, "Inaccurate" for outside 30%, and "Very Inaccurate" for indicating an incorrect direction), and (c) all results were presented on a summary at the end for further reflection. Initial interviews revealed that the display of the accuracy feedback was perceived as "too aggressive" and that the true value and source did not necessarily seem plausible to participants without explanation.</td></tr>
</table>

2



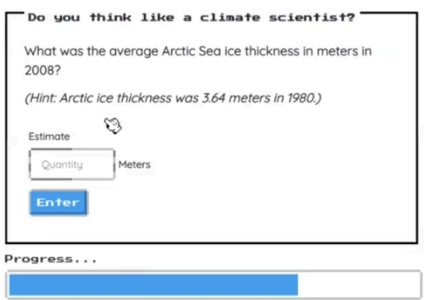


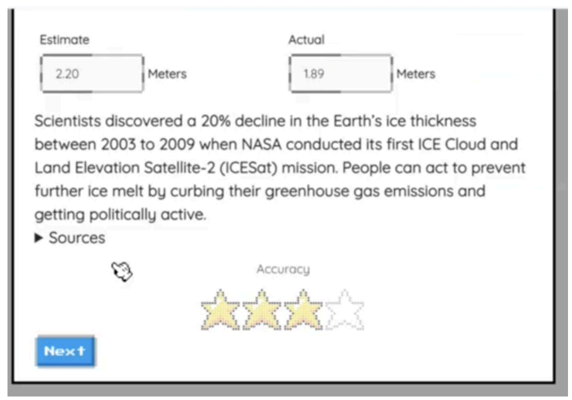


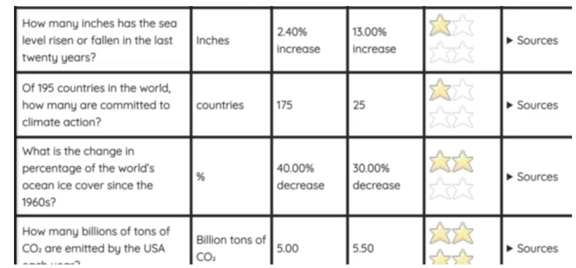
Fact	Units	Actual	Estimated	Accuracy	Source
What is the change in atmospheric levels of methane (a greenhouse gas) since 1750 until now?	%	151.00% increase	200.00% increase	Inaccurate	Sources
How many inches has the sea level risen or fallen in the last twenty years?	Inches	2.40% increase	10.00% increase	Inaccurate	Sources
What is the change in percentage of the world's ocean ice cover since the 1960s?	%	40.00% decrease	30.00% increase	Inaccurate	Sources
What was the average Arctic Sea ice thickness in meters in 2008?	Meters	1.89	3.00	Inaccurate	Sources
In 1850, there were approximately 150 glaciers present in Glacier	glaciers	25	50	Inaccurate	Sources

For the second iteration, main revisions involved (a) adding evidence and explanations for each true value, (b) including text prior to the intervention to improve tolerance for error, (c) revising the "look and feel" to be more game-like, and (d) decreasing emphasis on performance feedback by withholding the accuracy feedback until completion of the intervention. Interviews revealed that participants might benefit from reintroduction of the accuracy feedback with greater emphasis on positive aspects.

3







How many inches has the sea level risen or fallen in the last twenty years?	Inches	2.40% increase	13.00% increase		Sources
Of 195 countries in the world, how many are committed to climate action?	countries	175	25		Sources
What is the change in percentage of the world's ocean ice cover since the 1960s?	%	40.00% decrease	30.00% decrease		Sources
How many billions of tons of CO_2 are emitted by the USA	Billion tons of CO_2	5.00	5.50		Sources

The third iteration of the design mostly included revisions to how accuracy feedback was presented. Stars were introduced to represent accuracy with the same accuracy thresholds as presented in (1), (four stars for an estimate within 10% of the true value, three for within 20%, two for within 30%, 1 for outside of 30%, and no stars for indicating an incorrect direction). After this iteration, a fifth Star was eventually added so that one star was the minimum possible score.

Supplemental Materials

Appendix SA: Estimation Game Items

Items 1-6 drawn from Ranney & Clark (2016); Items 7-12 drawn from Thacker & Sinatra (2022). Instructions: *You will now estimate various quantities. After estimating each quantity, you will be shown the scientifically accepted value.*

Textual Description	Your Answer	Correct Value
1. What is the change in atmospheric levels of methane (a greenhouse gas) since 1750 until now?	% increase or decrease	151% increase
2. What is the change in percentage of the world's ocean ice cover since the 1960s?	% increase or decrease	40% decrease
3. According to observation data collected at Mauna Loa Observatory in Hawaii, what is the percent change in atmospheric CO_2 levels from 1959 (when observation began) to 2009?	% increase or decrease	22.6% increase
4. A 2010 article examines the 908 active researchers with at least 20 climate publications on Google Scholar. What percentage of them have stated that it is "very likely" that human-caused emissions are responsible for "most" of the "unequivocal" warming of the earth in the second half of the 20th century?	% of researchers	97.5%
5. In 1850, there were approximately 150 glaciers present in Glacier National Park. How many are present today?	# of glaciers	25
6. From 1850 to 2004, what is the percent change of volume of glaciers in the European Alps?	% increase or decrease	50% decrease
7. How many inches has the sea level risen or fallen in the last twenty years? *Hint: global sea levels rose by 1.0 inch between the years of 1900 and 1920, near the dawn of the industrial revolution.*	Inches (increase, decrease)	2.4"
8. What was the average Arctic Sea ice thickness in meters in 2008? *Hint: Arctic ice thickness was 3.64 meters in 1980.*	Meters	1.89m
9. How many square kilometers of Arctic ice cover was there in September of 2017? *Hint: September ice cover was 6.54 million sq kilometers before 2010, on average.*	Million km^2	4.8
10. How many billions of tons of CO_2 are emitted by the USA each year? *Hint: European Union, currently consisting of twenty eight countries, collectively emit 3.25 billion tons of CO_2 per year.*	Billion tons of CO_2	5
11. What is the change in average global temperature in the last 50 years? *Hint: the temperature increased by 0.7 °F between 1900 and 1950.*	Degrees F	1.62 °F
12. Of 195 countries in the world, how many are committed to climate action?	# of countries	175

Appendix SB: Survey Materials

Human Induced Climate Change Knowledge Measure

(HICCK; Lombardi, Sinatra, Nussbaum, 2013)

Below are statements about climate change. Rate the degree to which you think that climate scientists *agree with these statements. (Note that items with an asterisk directly relate to alternative conceptions about the causes of climate change summarized by Choi et al., 2010)*

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1	2	3	4	5

1. We cannot know about ancient climate change. (Reversed)
2. Earth's climate has probably changed little in the past. (Reversed)
3. Greenhouse gas levels are increasing in the atmosphere.
4. Earth's average temperature has increased over the past 100 years. This is evidence of climate change.
5. Average sea level is increasing. This is evidence of climate change.
6. Most of the world's glaciers are decreasing in size. This is evidence of climate change.
7. Most countries are committed to climate action.

Appendix SC

Table S1.
Descriptive Statistics Presenting Knowledge by Gain Group

	Total					**High Gain**					**Low Gain**				
	n	Mean	SD	Min	Max	n	Mean	SD	Min	Max	n	Mean	SD	Min	Max
Pre-Knowledge Mean	22	1.97	0.51	1.71	3.71	11	2.80	0.51	1.71	3.57	11	3.26	0.40	2.43	3.71
Post-Knowledge Mean	22	1.46	0.36	2.71	4.00	11	3.59	0.35	2.86	4.00	11	3.49	0.38	2.71	3.86
Gain (Post-Pre)	22	-0.51	0.35	-0.14	1.14	11	0.79	0.25	0.43	1.14	11	0.23	0.18	-0.14	0.43

Note Pretest and posttest knowledge variables represent mean scores ranging from 1 (minimum possible score on all seven items) to 5 (maximum possible score on all seven items).

Appendix SC
Cognitive Interview Protocol

Overview of Cognitive Lab Process

Introduction

- Greeting / establish rapport
- Read cognitive interview instructions aloud to the participant
- Practice exercises
- Model the think-aloud procedure
- Let the participant practice the think-aloud procedure
- Administer the interview or survey as follows:

PHASE I (Think aloud)

- Ask participant to read the questions aloud.
- Do <u>not</u> ask questions – allow the participant to think aloud as s/he answers the questions.
- If needed, use general prompts to encourage the think-aloud process (e.g., remember to tell me everything that you are thinking).
- Ask the participant if they are finished, or to indicate when they are finished.

PHASE II (Probing)

- Item-specific probing (see probes written on protocol)
- Goal is to understand why the participant responded the way s/he did.
- Designed to help you describe the process used by participants in responding, and identify any potential problems with the item.
- You may not need to use every probe – flexibility within structure!
- Another technique may be mirroring or summarizing what the respondent said to confirm that you understood.

Additional Probing for Estimation Game Items

- When participants have finished thinking aloud through all intervention items, ask all of the specific questions about usability:
    - Did you hit any problems or bugs?
    - What would you change about the look and feel of the game?
    - What would you change about the way information is presented in this game?
    - What was surprising?

Conclusion / Debriefing

- Answer participant questions
- THANK participant!

INTRODUCTION / PRACTICE EXERCISE

SAY:
Before we begin, please take a look at and sign the consent form. Take your time and feel free to let me know if you have any questions.

I am designing a survey for a large-scale study and I need your help to improve the quality of the questions that I ask before distributing the survey. Right now, I am trying out survey questions with people before sending the survey out to undergraduate students across the USA. This will help me identify problems with the questions that we created.

The way we will try this out is by having you read each question out loud, then answer the questions and tell us what you are thinking as you figure out your answer. This is called "thinking aloud."

Since people are not used to thinking aloud, I'd like to show you an example.

EXAMPLE 1 [Share Question With Participant]:
Suppose there was a set of multiple choice questions that asked: "In the past week, how often did you watch the following on television? (News, Drama, Sitcoms, Documentaries, Movies)." If someone asked me to think aloud while I was answering these questions, I would start by answering the questions, and then when I was finished, I would explain what I was thinking.

(Share screen [see Appendix] and read the instructions out loud and answer questions to yourself to illustrate thinking aloud)

In the past week, how often did you watch the following on television? (News, Drama, Sitcoms, Documentaries, Movies).

News.	[None]	[1-2 Days]	[3-4 Days]	[5-6 Days]	[7 Days]
Drama.	[None]	[1-2 Days]	[3-4 Days]	[5-6 Days]	[7 Days]
Sitcoms.	[None]	[1-2 Days]	[3-4 Days]	[5-6 Days]	[7 Days]
Documentaries.	[None]	[1-2 Days]	[3-4 Days]	[5-6 Days]	[7 Days]
Movies.	[None]	[1-2 Days]	[3-4 Days]	[5-6 Days]	[7 Days]

Now, you try it. In the past week, how often did you watch the following types of television?

EXAMPLE #2 [If they don't understand from example 1]

Suppose there was another question that asked: "How many minutes did you talk on the telephone in the last three days?" If someone asked me to think aloud while I was answering that, I would say:

Question: How many minutes did you talk on the telephone in the last three days?

SAY (for example): At work I usually spend about 30 minutes a day on the phone. Thirty minutes times three is 90 minutes. But, today, I had a few more calls than usual, so I probably spent about an hour on the phone. So, that brings the total to 2 hours. And that was just for business.

I was also in a meeting that was a telephone conference for another hour. But that's not really being on the phone, so I'm not going to count that because it's not just one person talking to another.

Also, does "the last three days" count today, or end yesterday? I'm not sure so I'll include today, even though today isn't over.

I also make some personal calls at work. Yesterday I called a friend at lunch and we talked for almost my whole lunch hour. That adds another hour, which makes three work hours. How much I talk at home is too hard to figure out, so I'm going to just kind of guess -- maybe 15 minutes each day. Times three, that's another 45 minutes. But, that seems kind of low, since I know I have a few long calls with my parents each week, so I'll raise that to an hour. So, three hours at work plus another hour at home -- that's four hours in the last three days. So, that would be 240 minutes.

Now, you try it. *How many minutes did you talk on the telephone in the last three days?*

Have participant run through the example question.

WHEN THE PARTICIPANT UNDERSTANDS (after the first or the second example):

That's is exactly what I mean by thinking aloud. Now we are ready to get started. Do you have any questions before we begin?

You will read the text in each item aloud, and you will think aloud as you answer it. Then you will read the next item and you will think aloud as you answer that one. This will continue until I stop and go back to ask you some questions about the items you already answered. Do you understand? OK, let's get started…

PHASE I
Examples of How to Elicit Think Aloud
(Try to ask each of these at least once during the interview)

Directions

1. The directions say (e.g.), "*You will now estimate various quantities. After estimating each quantity, you will be shown the scientifically accepted value.*" What does this mean to you?

Stem

2. What does the question stem (e.g.), "*We cannot know about ancient climate change.*" mean to you?

Answer Choices

3. How are you thinking about the answer choices?
4. Is there another option you would like to add to or remove from the answer choices?
5. Is the order of the options helpful in giving your answer?
6. Is there another rating scale you think would be more appropriate? Why?

Appropriate Prompts in Phase 1 to help the respondent think aloud include:

- *Go ahead and read the question, then answer it, we'll talk about it after you're finished.*
- *OK, I see, uh-huh, etc.*
- *Remember to **think aloud** as you answer the question.*
- *You're doing a great job **thinking aloud**.*
- *When you **think out loud**, it really helps us to understand how others approach these questions.*
- *Let me know when you are finished answering the question.*
- (If respondent struggles with answer) *Pretend this is a mail survey questionnaire you received, what would you do?*

PHASE II

After participants have "thought aloud" through cluster of items use general probes:

- OK, so what was your reasoning on that again? I just want to make sure I understand.
- *Can you say this question in your own words?*
- *What do you think this item is asking/measuring?*
- *How did you figure out your answer to this question?*
- *(If a word or phrase X is problematic) What do you think they mean by X?*
- *OK, so what was your reasoning on that again? I just want to make sure I understand.*
- *When you read this question, what did you think they were asking?*
- If you had to explain this question to a friend, what would you say they are getting at?
- So, what made you choose X, instead of say Y?

ADDITIONAL PROBING

Specific for Estimation Game Items

- When participants have finished thinking aloud through all intervention items, ask ALL of the following questions about usability:
    - Did you hit any problems or bugs?
    - What would you change about the look and feel of the game?
    - What would you change about the way information is presented in this game?
    - What was surprising?

What other kinds of content would you want to learn about with an experience like this?

Appendix SD: Codebook

How People Make Sense of Climate Change Numbers
Strategies and reactions involved when estimating and interpreting climate change quantities.

Estimation Strategies		Examples Quotes From Interviews
Mental Computation		
Benchmark Estimation (unspecified strategy)	Evidence that the individual used the given benchmark values without specifying what they specifically did with the benchmarks.	● I will use the hint
Benchmark Estimation + Extrapolation (trend thinking)	Evidence that the individual used the given benchmark information and beliefs about projected trends to estimate the unknown quantities, though the mathematical procedure may not be clear.	● So if we take into consideration that increase [given in the hint], and we use the same logic, so 0.7 per every 50 years.
Benchmark Estimation + Arithmetic (iteration/ proportional reasoning)	Evidence that the individual is explicitly using arithmetic. This could be repeating a given number over and over to estimate an unknown number, or multiplying, dividing, rescaling a number to obtain their estimate.	● I'll put 1.5% because it's half of 3% ● I'm going to put two inches will say we doubled it. ● So we'll add another third to that so 1.24 plus point three... I'll say is 1.57
Flexible Rounding	Evidence that the individual has rounded a given number to make mental computation easier.	● I [rounded to] 194 because [halving] 195 will result in a point five calculation, and there's not really half a country, so I just rounded down because, you know, down is less
Tolerance for Error		
	Evidence that the individual feels comfortable with imprecision. (E.g., "My estimate wasn't perfect, but close enough to be reasonable," or "wow, my guess was good. It was pretty close").	● Oh, my gosh that's not bad. 1.6 in comparison to five it's bad, but like in comparison [to my estimate] that's not bad at all. I'm okay with that.
Prior Knowledge		
	The individual references information from their prior learning experiences OR personal experiences to estimate or make sense of unknown quantities (e.g., "I saw a documentary once that said that methane has basically doubled in the last 50 years… so I'll double this number" or "It has been getting smoggier [where I live], so I think that CO2 has increased…").	● [based on] the documentaries I've seen in my life, they been saying things like [temperature] generally increases ● I know because I researched a little bit about it ● I heard about that somewhere, like on radio or podcast or Facebook ad, something
Positive / Negative Reactions to Accuracy Feedback		
	Evidence that the individual is interpreting the accuracy feedback when provided with the true value (e.g., "I am not doing very well." or "Wow,	● Look at that. Like that is perfect." ● Yeah I got that pretty much right. Overwhelming majority, 9 out of 10 ● Oh, my gosh what did I get this

	that is the third red in a row. I am not doing well”).	wrong… What okay, bye.” ● Yeah, this is like ‘oh my gosh, she’s so terrible’
Wild Guess	Evidence of wild guessing. Seems to be “pulling a number out of nowhere” without much consideration or justification for where the number came from. No context whatsoever given for the estimate.	● I would just guess blindly [in this case] because I don't think the number indicates like is anywhere close. ● This is just a pure guess ● I would totally guess here because I don't know anything about this

Note. For this preliminary study, subcodes were consolidated into parent codes (e.g., “Mental computation” represents occurrences of mental iteration, proportional reasoning, *and* flexible rounding).