Question Based Learning: A Pedagogical Approach for Improving Hypothesis Generation in Active Learning

by

Grace K Wallace

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2021 by the
Graduate Supervisory Committee:

Nicholas Duran, Chair
Kelsey Lucca
Zachary Horne

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

Science education faces a distinct challenge in the transition to active learning: how can teachers ensure students reach accurate understandings during the exploration and self-discovery phase of a lesson? Research in hypothesis generation demonstrates human's vulnerabilities to specific biases based on prior knowledge, selective memory retrieval, and failure to consider alternative explanations. This is further complicated in science education, where content standards are abstract. As such, it is imperative to implement a proactive intervention to curb misconceptions from forming during active learning in science lessons. In this work, a new a model of instruction, Question-Based Learning (QBL) is designed and tested against current learning paradigms. The study aims to investigate whether providing constraint-seeking questions is an effective intervention leading to improved mastery of learning targets during active learning. Participants were randomly assigned to one of three conditions to learn a scientific concept: a blended learning condition, a guided-inquiry condition, or a QBL condition. Mastery was measured at the end of the task using a 12-question assessment. The same measure was also administered one week after subjects completed the study to see whether delayed recall significantly differs between condition groups. Results indicate the QBL model is at least as effective two existing forms of pedagogy at teaching a scientific principle, increasing depth of knowledge regarding that scientific principle, and sustaining knowledge over time.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS

Page

LIST OF TABLES

Figure Page

LIST OF FIGURES

Figure Page

Question Based Learning: A Pedagogical Approach to Improving Hypothesis Generation in Active Learning

The traditional rote lecture style learning that once dominated K-12 school systems now only comprises about 5- to- 10 minutes of most lesson plans (Chi, 2009; Chi & Wily, 2014; Levy et al., 2013). The remaining and majority of classroom time (40 to-60 minutes) involves inquiry- and play- based learning activities, in which students are engaged in open-ended and kinesthetic learning environments without direct instruction from a teacher (Levy et al., 2013). Such activities exhibit the pedagogical shift from passive learning, where students receive and memorize information from a source such as a teacher or text, into active learning environments that allow students to self-direct their learning and discovery of new concepts (Chi & Wiley, 2014).

Active learning encompasses a broad range of definitions, teaching techniques, and lesson structures, however, three major core components exist across varying content areas. First, the pace of the lesson should be student-directed rather than teacher-directed (Chi, 2009; Chi & Wily, 2014; Levy et al., 2013; Markant et al., 2016). Self-pacing allows for variance in attention processing, metacognition, and adaptive studying techniques; stated plainly, it allows for differentiation in the size of information chunks, speed of input, and reevaluation of prior knowledge based on individual student need (Chi & Wiley, 2014; Markant, 2018, Markant et al., 2016; Preston & Eichenbaum, 2013). Secondly, all active learning is constructive, meaning students self-discover the desired learning target through observation and experimentation as opposed to being told

the learning target through direct instruction (Bonawitz et al., 2011; Chi & Wiley, 2014; Engel, 2011; Levy et al, 2013). Thirdly, active learning is meant to be collaborative; this provides students an opportunity to revise understandings through the assistance of their peers (Chi & Wiley, 2014; Levy et al, 2013).

Work across education, psychology, and cognitive sciences indicates these components make active learning largely more efficacious than passive learning. Active learning models increase memory retention by 5-10 percent and promote deeper conceptual understandings of rigorous content (Markant et al., 2016). When utilized in undergraduate STEM courses, these benefits led to a significant 12 percent reduction in failure rates and 0.11 point improvement in average GPA (Freeman et al., 2014; Levy et al., 2013). These numbers are even more substantial for underrepresented student populations – highlighting active learning as a viable solution for both producing graduates and diversifying the STEM workforce (Freeman et al., 2014).

In classrooms, however, teachers note a major pitfall of active learning: it assumes the cognitive capacity to correctly discover the desired learning target. Herein represents the existing gap for practicing educators – how do we ensure students reach accurate understandings during self-guided exploration (Bonawitz et al., 2011; Chi & Wiley, 2014; Engel, 2011; Levy et al, 2013; Markant et al., 2016; Markant, 2018)?

This process, recognized in cognitive science as hypothesis generation, is the mechanism by which humans form the explanations and conclusions regarding witnessed phenomena that then drive further reasoning, prediction, and exploration (Markant et al, 2016). But these explanations and conclusions are prone to specific biases based on prior

knowledge, selective memory retrieval, and alternative explanations (Ruggeri et al., 2019). Science education is especially disposed to such issues due to its abstract and environmentally inconsistent content. For example, salt water has a lower freezing point than regular water, meaning liquified salt water will be colder in temperature than frozen water. This goes against our natural and observable assumptions because we are inclined to believe that frozen substances are at lower temperatures than liquid substances. In other words, science education disproportionately suffers from hypothesis generation bias because the content is not always constant or conclusive and, furthermore, the content requires students to reach hypotheses not always observable when judged by the human eye (Chi, 2009; Chi & Wily, 2014; Levy et al., 2013).

Such biases pose two potential issues for science educators: 1) students fail to generate a hypothesis because the task is too cognitively demanding (i.e. students fail to account for the role of gravity when calculating the speed of a free-falling object because gravity is not "visible"), or 2) students generate an incorrect hypothesis and, as a result, form a scientific misconception (i.e. students assume the acceleration of gravity is dependent on object size and mass). In both cases, the student fails to reach the desired learning target, leading to decreased learning outcomes (Bonawitz et al., 2011; Chi & Wiley, 2014; Engel, 2011; Larsen et al., 2019; Levy et al, 2013; Markant, 2018). If active learning is to be an effective and sustainable methodology for improving science education, we must work to close this gap.

Our first aim in this work, therefore, is to expand upon a strategic pedagogical model of science instruction that relies on question-asking to improve hypothesis

generation. Though simplistic in nature, this model, which we titled question-based learning (QBL), mimics the learning methods used throughout early development: observation, experimentation, and modification (Lucca & Wilborn, 2018; Ronfard et al., 2018; Stahl & Feigenson, 2015). In QBL, teachers will provide viable questions for a variety of potential hypotheses that students self-select to autonomously explore – letting students formulate, investigate, and amend beliefs as part of the learning process. This procedure differs from current practice in that it both a) requires students to ask questions rather than answer questions, more closely aligning with cognitive mechanisms used in early childhood, and b) allows students to entertain and discern between a multiple of plausible hypotheses, a fundamental aspect of real-world problem solving (Engel, 2008; Lucca & Wilborn, 2018; Ronfard et al., 2018; Stahl & Feigenson, 2015).

Secondly, we will assess how our QBL model compares to existing learning paradigms — determining whether its effectiveness varies as a function of content rigor and across time. Subjects will complete one learning task for a scientific principle (kinetic molecular theory taken from general secondary science curriculum). Subjects will either be directly taught the learning target (blended learning condition), will watch real-world demonstrations and be asked to answer a series of scaffolded questions (guided-inquiry condition), or will be given various questions to select for exploration (QBL condition). After the learning task, subjects will be assessed on an immediate and delayed basis. We hypothesize QBL will increase content mastery on rigorous, application-based assessment and this increase will be prolonged over time – extending QBL as a potential future pedagogical direction for science education.

Our study seeks to expand upon past research demonstrating question-asking as a viable pedagogical tool in science education. Previous work suggests students high in self-regulatory processes, such as planning, metacognitive monitoring, and revision, are able to excel in student-directed and open-ended exploratory learning environments largely because such students are able to recognize their current levels of understanding and ask appropriate follow-up questions in light of that understanding (Azevedo, 2009; Azevedo, 2005). Investigations regarding tutor-peer interactions furthermore support this question-asking framework; learners exhibit higher knowledge gains when tutors utilize questioning rather than telling as a pedagogical tool (Roscoe, 2014; Roscoe & Chi, 2007). Similar findings are also demonstrated when the tutor is removed and instead replaced with computer simulated questions and explanations for learners to explore and gain knowledge in active learning environments (Graesser & McNamara, 2005). Our novel contribution in this work is to consolidate these findings into a pedagogical mechanism reliant upon question-asking that K-12 science teachers could then utilize as a lesson planning framework.

Though question-asking has long been the learning mechanism for children, schools often flip this methodology – instead asking students to answer questions rather than explore their own curiosities (Bonawitz et al., 2011; Engel, 2011; Levy et al., 2013). This approach, however, fails to build skills involved in hypothesis generation because it lacks the opportunity for exploration and revision of incorrect theories. And though in the short-term, foregoing hypothesis generation allows educators to quickly disseminate information to large groups of students, it also ignores the type of critical thinking needed

in applied settings (Bonawitz et al., 2011; Engel, 2011; Freeman et al., 2014). Our work, therefore, will extend upon current pedagogical interventions to better scaffold the cognitive demand of hypothesis generation, expanding our ability to implement active learning and produce the necessary problem solvers for today's 21st Century world.

Literature Review

The pedagogical shift to active learning already cemented its role and value for improving science education as demonstrated by its advances in memory retention, conceptual understanding, and tangible academic outcomes. Questions still remain, however, regarding a "second-generation approach" to research, focusing on the specific activities, interventions, behaviors, etc. to maximize learning under such conditions (Bonawitz et al., 2011; Engel, 2011; Freeman et al., 2014, Levy et al., 2013; Markant, 2018). This how-to focus represents the current missing link for educational professionals: through what type of pedagogical instruction do we best maximize hypothesis generation, accurate understandings, and knowledge retention? Our study is intended to be part of that second-generation research — designing and testing a pedagogical model for science education to improve active learning outcomes.

Issues in Hypothesis Generation

In science courses, students typically engage in active learning by using observed phenomena to discover scientific principles (Chi, 2009; Chi & Wiley, 2014; Levy et al., 2013). For example, two objects of different masses dropped at the same time fall at the same rate (the observed phenomena), demonstrating the conceptual understanding that the acceleration of gravity is constant (the scientific principle). Again, this process –

hypothesis generation – forms the explanations and conclusions for witnessed events which then drive further reasoning, prediction, and exploration (Markant et al., 2016; Markant, 2018).

Hypothesis generation, however, is a highly demanding cognitive task and is susceptible to numerous biases including memory retrieval, prior knowledge, failure to consider alternatives, and confirmatory explanations (Markant, 2018). The potential failure to ascertain accurate theoretical knowledge often leads to broad misconceptions and an inability to produce scientific knowledge – drastically reducing learning outcomes (Levy et al., 2013; Markant, 2018). And, from previous work in scientific belief revision, we know children are unlikely to change misconceptions even if faced with anomalous physical evidence or subsequent direct instruction (Larsen et al., 2019). This highlights the difficult task of balancing the benefits of active learning with the issues of novel hypothesis generation; though more engaged and able to recall inquiry- and play- based activities, students often fail to accurately and fully comprehend the conceptual component of such lessons.

For example, take the scientific principle mentioned above: the acceleration of gravity is constant and, therefore, all objects in a free-fall will have the same acceleration regardless of object size. Students low in prior knowledge might fail to identify the role gravity as the downward force acting on free-falling objects. Instead, students might assume objects fall because there is nothing holding the object up and, therefore, the heavier an object, the faster the fall. These students are likely unaware of or unable to apply Newton's 1st Law of Motion (an object at rest will stay at rest unless acted upon by

some outside force), and are unable to reach the theoretical conclusion that a free-falling object will not actually "fall" unless acted upon by a downward force (gravity).

Other students, alternatively, might be aware of the role gravity plays in forcing the objects downward, but could assume gravitational force is dependent on the object's size and mass. Such students might fall victim to confirmatory analysis if they were to test their theory on objects that experience vastly different forces of air resistance (i.e. a poster board versus a bowling ball). The poster board would fall significantly slower than the bowling ball, "confirming" their belief that the bowling ball – due to a larger mass – experiences a larger downward gravitational acceleration. Both sets of students would reach the same incorrect hypothesis: heavier objects fall faster and, therefore, have a higher acceleration. These hypotheses, however, would be rooted in two completely different conceptual understandings.

This example exemplifies the difficulties surrounding hypothesis generation and using hypothesis generation as a learning component in classrooms. Educators cite these limitations as a major obstacle to implementing active learning (Levy et al, 2013). Experimental work additionally demonstrates that individuals perform better when demand for hypothesis generation is reduced during active learning tasks (Markant, 2018). It is, therefore, imperative to design pedagogical interventions for ameliorating hypothesis generation.

Current Pedagogical Models

Currently, lesson plans typically utilize pedagogical paradigms to circumvent the hypothesis generation component of active learning. Two such models are blended

instruction and guided-inquiry instruction (Chi, 2009; Chi & Wiley, 2014; Levy et al., 2013). In the former, teachers use a combination of passive and active learning to reduce the cognitive requirement on students (Chi, 2009). In the latter – guided-inquiry instruction – students are given a set of directions to follow and scaffolded questions to answer that guide the learner toward a specific understanding (Chi & Wiley, 2014; Levy et al., 2013). The biggest and most significant difference between these two paradigms is the learning expectation on students; in blended learning, students are explicitly taught the learning target whereas students are expected to generate the learning target when undergoing guided-inquiry learning. Similarities exist, however, in that blended learning and guided-inquiry instruction work by reducing the hypothesis space to a single conclusion (Bonawitz et al., 2011; Fox et al., 2019; Levy et al., 2013).

Used early-on in active learning methodology, blended learning is the least cognitively rigorous model as it eliminates the process of hypothesis generation entirely – giving students the scientific principle outright through direct instruction (Chi, 2009; Chi & Wiley, 2014; Fox et al., 2019; Levy et al., 2013). For example, students might be explicitly told the learning target (i.e. the acceleration due to gravity of a free-falling object is constant) and asked to prove this theory (i.e. being given objects of various masses to drop and observe). Though an effective way to disseminate accurate information and avoid misconceptions, blended learning is widely criticized for lacking student engagement opportunities. Because students already know the answer, there is no motivating information gap to fill (Litman, 2005; Lowenstein, 1994; Pluck et al., 2011). Psychology has long recognized the importance of curiosity as an effective mechanism

by which the learning and storage of information occurs (Berlyne, 1954; Lowenstein, 1994); educational pedagogy that stymies curiosity, therefore, holds little value in promoting learning outcomes beyond that of simply ensuring student understanding is accurate (Bonawitz et al., 2011; Engel, 2011; Litman, 2005).

Guided-inquiry instruction, on the other hand, was developed as an improved alternative model. It requires increased cognitive reasoning by asking students to answer questions using observation and experimentation as a way to narrow the hypothesis space toward the correct scientific understanding (Chi & Wiley, 2014; Levy et al., 2013). Using the same example regarding free-falling objects, students might receive the objects of various masses, observe what happens when they drop the different objects at the same time, and answer a series of questions to clarify the role of gravity and its relationship to the object's acceleration. This type of pedagogy is supported by cognitive work which concludes one's ability to learn a new concept without prior direct instruction is dependent on how the individual is guided through the learning environment (Markant, 2018). Studies show scaffolded supervision during the hypothesis generation phase can significantly improve learning outcomes (Markant, 2018). Furthermore, providing students with a scaffolded question sequence to generate a specific hypothesis does, in fact, increase the perseverance and the exploratory actions taken by children when engaged in a learning task (Jean et al., 2018).

Despite the augmented demand, however, guided-inquiry instruction still follows the oft-criticized formulaic educational model in which informants (teachers) ask questions while students answer (Bonawitz et al., 2011; Engel, 2011; Liquin &

Lombrozo, 2017). Those answers, furthermore, are only directed to guide the learner to a particular (accurate) solution, reducing the potential for student curiosity to drive exploration, investigation, and learning (Bonawitz et al., 2011; Engel, 2011; Liquin & Lombrozo, 2017). At first glance, it seems counter-intuitive to suggest limiting the hypothesis space to only one option is detrimental to learning as educators are pointing students to the correct answer. And if judging learning by performance on a standardized measure designed to reward the "drill-and-kill" approach - a tactic in which students are taught through a systematic repetition of concepts - both blended learning and guided inquiry will likely yield quality results (Bonawitz et al., 2011; Engel, 2011). But this measurable success exemplifies the dilemma of inductive bias: approaches currently used by educators to deliver pedagogical instruction "necessarily limit the range of hypotheses for student consideration to promote rapid and efficient learning of desired material" (Bonawitz et al., 2011, p. 324).

For example, when children were given a novel toy with multiple functionalities, they were more likely to focus exclusively on target function when the function was directly modeled or when children were asked what happened when that specific aspect of the toy was manipulated. When direct instruction was not provided, however, children engaged in broad exploration of the toy's many novel functions (Bonawitz et al., 2011). In this instance, children learned toy functionality not because they were directly taught or asked to learn toy functionality, but rather, because their curiosity drove such learning. This type of uninhibited exploration, though less direct, pushes learners to grapple with unknowns and infer conclusions as if engaging in problem-solving.

Real world problem-solving – the kind required in applied fields such as the sciences – is rarely limited to a singular premise. Instead, it entails crafting a multitude of hypotheses, proactively eliminating misconceptions through investigation, and condensing evidence-based results into theoretical understandings (Bonawitz et al., 2011; Engel, 2011; Jean et al., 2018; Liquin & Lombrozo, 2017). It demands our brains excel in hypothesis generation. In this sense, current pedagogical models lower the quality of learners' critical thinking; they simply directly or indirectly pursue one plausible conclusion with little regard for engaging student curiosity. Therefore, while blended learning and guided-inquiry do provide a base paradigm for active learning, they fall short of correcting a root issue facing students in science fields: hypothesis generation.

Proposed Model: Question-Based Learning

Rather than eliminating hypothesis generation from active learning in science education, our pedagogical model, QBL, embraces it. Our model is designed such that students use their cognitive resources to ask, explore, and analyze questions instead of answering them. Question-asking is a tool employed by children far before language develops – with toddlers using pointing gestures to elicit needed information – and continues to be our greatest device for knowledge acquisition through much of childhood (Lucca & Wilborn, 2018; Stahl & Feigeson, 2015). By pushing students to ask the questions they are interested in, we activate those same mechanisms already in our cognitive toolbox to engage curiosity and drive learning.

Secondly, it also allows individuals to investigate potential hypotheses based on their current predictions and understandings. This necessarily differentiates the

hypothesis space based on prior knowledge and curiosity, which could encourage individuals to address potential misconceptions during the learning process (Fox et al., 2019; Larsen et al., 2019; Markant, 2018). Advocates for an inquiry based approach to teaching and learning suggest self-regulation is a necessary developmental skill by which students are able to plan, monitor, and control their cognition, motivations, and behaviors for successful knowledge building (Azevedo, 2009; Azevedo, 2005). Our QBL model is designed such that students are required to engage in self-regulated learning to move through the learning task; again, pushing students to not only engage in knowledge-building, but to do so in a way that is relevant to their current understandings and any formed misconceptions.

Previous work furthermore indicates this movement through the hypothesis space — referred to as self-monitoring and knowledge-building — is a crucial component in which tutors can increase student depth of knowledge (Roscoe, 2014). During individualized student-tutor interactions, the tutor's ability to avoid a *knowledge-telling bias* — instead building student knowledge through reasoning and questioning — leads to significantly greater learning outcomes for students (Roscoe, 2014; Roscoe & Chi, 2007). Our QBL model works similarly, albeit without the tutor, in that it seeks to build knowledge through a reasoning and questioning structure that can be scaffolded to students' current level of understandings.

Finally, our question-based learning model also recognizes the need for scaffolding the hypothesis space. Leaving students in an open-ended environment without direction is too challenging of a cognitive task (Markant,2018; Ruggeri et al.,

2017). To reduce this demand, we recommend teachers provide potential initial and follow-up questions throughout the task. Students would then self-select which of these questions to investigate. Previous research demonstrates adolescents are able to recognize which questions provide purposeful information gain, even if they are unable to dictate such questions on their own (Ruggeri et al., 2017; Ronfard et al., 2018). Work with computer-based learning modules suggest this differentiated level of choice for building questions and explanations of knowledge also leads to increased learning outcomes (Graesser & McNamara, 2005). It is our belief, therefore, providing students with a set of reasonable questions to explore within a self-directed learning environment will sufficiently scaffold the hypothesis space to a manageable cognitive load while still exploring a multitude of theories as students test, retest, and revise understandings.

The QBL model brings an innovative – and more importantly, easily implementable – approach to improving science education. At its core, it relies on fundamental aspects of active learning (self-direction and exploration) already familiar to educators and proven to be an effective form of pedagogy. But it enhances this methodology by reducing the barriers in hypothesis generation while pushing students to think critically as objective problem solvers.

Prior Research

Prior to our current work in which we test a QBL model against other existing forms of pedagogy, we conducted two observational studies to establish plausible scientific misconceptions as well as the investigation, design and piloting of a QBL task for a single scientific principle.

Preliminary Study 1

In an observational study conducted with 250 adults from Amazon's Mechanical Turk using Qualtrics, we examined subject beliefs regarding 5 different scientific principles before and after engaging in an active learning activity using a guided-inquiry instructional model.

The goal was to see if subjects were able to accurately identify learning targets based on scientific stimuli and if they were willing to change their conclusions regarding those learning targets at the end of the study. The study was purely descriptive so no manipulations or interventions were used, and no hypothesis was made for how people would perform on these learning tasks.

At the start of the study, subjects were randomly assigned one of five different scientific stimuli. We selected stimuli from topics in secondary science known for misconceptions. These included: free falling objects, floatation, temperature, velocity, and kinetic gas laws. The desired learning targets for each principle were taken from the national Next Generation Science Standards (NGSS Lead States, 2013).

Subjects were first asked to predict which scientific principle was most accurate from four plausible conclusions to generate pretest data regarding what subjects know and what misconceptions existed at the start of the study. Subjects were then presented with a stimuli specific story problem – such as predicting whether a heavy or light object falls to the ground faster and why. This portion of the survey primarily served as a way to engage subjects and determine if subjects could identify some of the basic measures in the scientific method, which is frequently used in K-12 science classrooms.

We used multiple choice questions with four possible answer options for each of the following – could subjects identify the scientific question being asked in the story problem (which object – a heavy or light – falls with faster acceleration when dropped at the same time), what factor is being manipulated for the two objects (mass), and what factor is being measured as a result (object's speed of fall). Finally, subjects made a prediction about what would happen in the scenario using a drop-down box – subjects could select the heavy object would fall fastest, the light object would fall fastest, or both objects hit the ground at the same time.

After engaging subjects, we then presented one of two randomized videos during an active learning task. The videos were selected to allow subjects to visualize the scientific principles for each stimuli – our videos for the free fall stimuli included a bowling ball and feather being dropped in a vacuum chamber and a hammer and feather being dropped on the moon. The videos demonstrate that both objects – regardless of mass – hit the ground at the same time.

After the videos, we asked subjects to complete a true or false drag-and-drop sorting task with 8 statements. The statements were based on an underlying understanding subjects could use to generate accurate hypotheses about the scientific principle causing both objects to hit the ground at the same time. Though no formal hypotheses or experimental designs were conducted, we were interested at a purely observational level whether subjects with higher accuracy on the true or false sorting question would also be more likely to select the correct conclusion, meaning the most accurate statement for the relevant scientific principle.

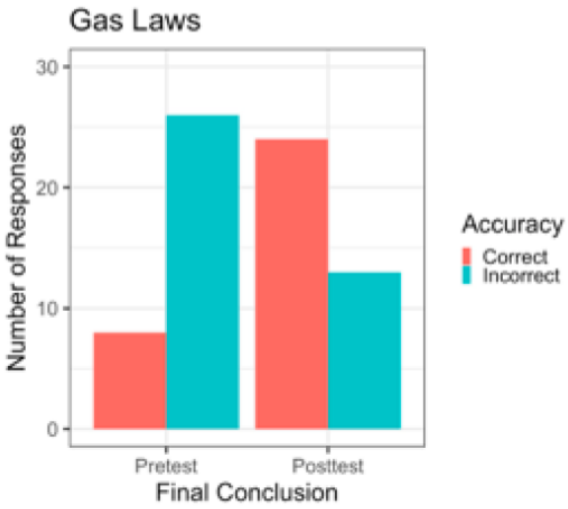
So – for example – a subject might observe the bowling ball as having more momentum than the feather because the video clearly depicts the bowling ball hitting the ground harder than the feather. This points to the fact that – although both objects fall at the same rate – the bowling ball falls with a greater force of gravity. From here, they might reason higher momentum does not correlate with faster acceleration and therefore, objects with greater force are not necessarily traveling at greater speeds. By this reasoning, it makes sense both objects could have the same acceleration since we observe them falling at the same rate.

Finally, subjects were then asked to return to the same question they had answered at the beginning of the survey. In our example, this is – again – the understanding that the acceleration of gravity is constant on all objects in a free fall regardless of their weight. This serves as an observational posttest measure to see whether subjects were able to identify the correct conclusion by the end of the task and whether their beliefs had changed from the beginning of the study.

Our analysis only examined general, descriptive data to identify how subjects think about these specific scientific stimuli and how those opinions might change after initial predictions. Our descriptive results are outlined below - again, we want to make clear that this is purely observational data as there was no control condition. See our results in Figure 1.

Figure 1

Findings of Preliminary Study 1



Note: A bar graph demonstrating the number of correct and incorrect responses for an accurate scientific conclusion from pretest versus posttest for the gas laws stimulus.

As indicated by the growth of the red bars from the pretest to posttest results, the number of subjects selecting the correct scientific conclusion at the end of the study increased. We are unable to say without a control condition whether this is due to the learning task or some other variable, however, it does appear to demonstrate subjects could be willing to change their minds from prediction to conclusion. Furthermore, subjects scoring higher on the true or false sorting activity also selected the correct scientific conclusion more often at the end of the study. Again, whether this result is due to the learning task is unknown, but it does give some plausibility to the idea that individuals higher in core knowledge were able to generate more correct hypotheses within our study.

With this observational data, we felt confident moving forward in this hypothesis generation research and specifically with these scientific stimuli as it does appear subjects have misconceptions prior to the learning task as seen by the significantly larger quantity of incorrect scientific beliefs at the pretest measure, and subjects could be willing to change these beliefs demonstrated by the shift in number of correct responses at the posttest measure. Additionally, and beneficial for our QBL model, it gave us some insight as to what misconceptions exist; for example, subjects most often made the hypothesis that gas volume is unrelated to temperature. This information was informative when building example questions for our QBL model to address commonly held, but incorrect, beliefs.

Preliminary Study 2

We then began to build a QBL pilot model using learning targets from the scientific principle kinetic gas laws. We specifically selected this principle because it was one of the scientific stimuli identified as high in both misconceptions at the pretest and corrected understandings at the end of the learning task in our first preliminary study.

We conducted a second observational study with 100 adult participants from Amazon's Mechanical Turk. We had multiple objectives in performing this pilot. First, we wanted to ensure the Qualtrics platform would aptly support our QBL model. Secondly, we were curious if subjects would voluntarily participate in the QBL design. We were specifically interested in how many questions subjects would choose to explore to assess whether it would be necessary to ask subjects to investigate at least two questions from the main screen in future studies.

To assess knowledge and learning at a pre- and posttest measure, we first designed a 12-question quiz for kinetic gas laws modeled after essential learning targets taken from the Next Generation Science Standards. All questions were multiple choice. The assessment was divided into three distinct categories of rigor – low, medium, and high – based on the first three levels of the Depth of Knowledge chart: Recall, Application, and Strategic Thinking (Webb, 1997). The four low rigor questions asked subjects to identify basic concepts regarding gases under normal behaviors such as w*hat is the definition of a gas*? Questions at the medium rigor section required subjects to determine how gas behavior might change under varying conditions (i.e. *if the temperature of a gas decreases, what happens to the amount of space the gas takes up?*). The high rigor section then asked subjects to relate visible behavior changes to particle behavior at an abstract level (i.e. *how do gas particles behave differently in cold temperature environments*).

After completing the pretest, subjects were then presented with a real-world story problem to promote understanding and engagement before beginning the learning task. Subjects were told they would select questions to explore that could potentially help them to explain the phenomena observed in the story problem. All subjects saw the main question screen, which included four potential questions for investigation. Subjects were told there was no specific correct answer, but to pick the question they were most curious to know in relation to the story problem. The questions were generated using common subject misconceptions and beliefs regarding gas particles and gas behavior in past research (Jauhariyah et al., 2018; Mayer, 2011; Nakhleh, 1992).

Participants were then directed to a video demonstration of the selected question – a timing feature was used to encourage participants to watch the video in its entirety. After the video, subjects were presented with the option to either ask a follow-up question related to the first question (i.e. *does changing the size and shape of a container change the amount of gas particles inside?*) or to return to the main questions screen. If subjects opted to engage with the follow-up question, they were directed to an additional video demonstration before returning to the main screen. Once back to the main screen with the original four questions, participants were free to select an additional exploration pathway or take the posttest and end the survey.

Again, we did not perform any specific analyses with this data; we did, however, determine that the majority of subjects voluntarily selected to explore an average of 2.7 questions, indicating a question requirement was likely not necessary for future studies. Additionally, the majority of participants (n = 77) indicated they were able to complete the survey without issue in the design, which allowed us to assume the directions and survey flow was conducive to the online learning task.

Current Work

Considering our observations from the above preliminary studies, we were confident in continuing our research on the QBL model. As such, our current work tested the QBL model design from Preliminary Study 2 against both a blended learning and a guided-inquiry learning condition. Our goal was to determine whether participants in the QBL model performed significantly better than those in the other conditions on the final

posttest, indicating the QBL model is at least as effective as two other existing learning models.

Because current research shows there is little difference in learning outcomes between passive and active learning models for less rigorous understandings, we anticipated our model would be most effective at improving outcomes for rigorous, application-based assessment (Levy et al., 2013; Markant, 2018; Markant et al, 2016; Ruggeri et al., 2019). Furthermore, we believed this difference would increase over time as with other active learning models in which increased subject accountability and self-direction improved prolonged learning outcomes (Chi & Wiley, 2014; Levy et al., 2013; Markant, 2018; Ruggeri et al., 2019). Such findings would provide some evidence our QBL model could be a valid and effective form of pedagogy for active learning instruction in science classrooms. Our research question, therefore, was: Did subjects in the QBL condition learn and retain more and deeper content knowledge than subjects in either the blended learning or guided-inquiry conditions?

In Phase 1 of our current work, we used a repeated measures pretest/posttest experimental design to investigate between-subject condition differences in knowledge gain with subjects randomly assigned to either a blended learning, guided-inquiry learning, or QBL condition. Phase 2 then examined between-subject condition differences for within-subject knowledge decay over time by asking subjects to complete a delayed posttest measure one week after Phase 1 concluded.

We made three hypotheses for these studies: 1) Subjects in the QBL condition will score higher on the immediate posttest than those in the blended learning or guided-

inquiry conditions, 2) There will be an interaction effect for condition and assessment rigor such that subjects in the QBL model will demonstrate significantly higher scores on the medium and high rigor questions of the assessment at the posttest measure than those in the blended learning or guided-inquiry learning conditions, and 3) There will be an additional interaction between condition and time such that subjects in the QBL condition will retain significantly more knowledge over time than those in the blended learning or guided-inquiry learning conditions as assessed by the week-long delayed posttest measure.

Participants

Due to varying perspectives on appropriate ways to conduct power analyses for mixed model designs, we opted to utilize both a paired and two-sample t-test power analysis to determine sample size since there is both a between-subjects factor of condition and a within-subjects factor of time. We used power = 0.80 and $d = 0.50$ based off of prior work from educational psychology with a high effect size (d=0.83) - assuming our effect size would be below this value - and the conventional significance level of 0.05 (Rorher et al., 2020). Using the pwr.t.test function in R-programming, we determined a paired-samples t-test would require 27 participants per group and a two-sample t-test would require 64 participants per group. We then selected 50 participants per group as a value between both test options, multiplied by 1.3 to account for participant drop out, to bring our full desired starting sample size to 195 total participants or 65 participants per condition.

We recruited 195 adult subjects from Amazon's Mechanical Turk. 190 subjects completed the entirety of Phase 1, which included the pretest measure, condition-based learning task, and the immediate posttest measure. The subjects were randomly assigned to one of three conditions: *blended learning condition* (*n*=66), *guided-inquiry learning condition* (*n*=65), or *QBL condition* (*n*=64). The majority of our subjects were first-language English speakers (*n*=162) from the United States (n=179) and were college graduates (*n*=133). 98 participants were male and 89 were female, with *n*=2 additional non-binary participants and *n*=1 participant identifying as other.

Subjects were paid for their participation. Subjects (*n*=189) that passed an attention check embedded in the immediate posttest were paid additional compensation for participation in Phase 2, which included our delayed posttest measure to analyze the effect of time on knowledge retention across groups. Subjects that did not pass the attention check (*n*=1) were not asked to participate in Phase 2.

139 returning participants then completed the delayed posttest in Phase 2 with a roughly equivalent number of returning participants from the condition groups randomly assigned in Phase 1 (blended learning condition, *n*=45; guided learning condition, *n* = 49; QBL condition, *n* = 45). Subject responses from Phase 1 and Phase 2 were matched using two different self-reported authentication factors including subject's Amazon Mechanical Turk identification number and subject birth year. All subject responses from Phase 1 and Phase 2 were able to be matched and, therefore, there was no additional data loss beyond attrition from Phase 1 to Phase 2.

Although we fell slightly short of the original goal to have to 50 participants in both Phase 1 and Phase 2 for each condition, all conditions were at or above halfway between our required sample size for a paired-samples t-test and two-sample t-test. As such, we believe our sample size was appropriate.

Materials and Methods

The study was longitudinal in nature and measured subject understanding of a given learning target at three separate time points - a pretest prior to any learning task, a posttest immediately following the learning task, and a delayed posttest one week after the learning task. For clarity purposes, we refer to the data collection at the pretest and immediate posttest as Phase 1 and the delayed posttest as Phase 2.

The same learning target was used for all subjects in Phase 1 and Phase 2. Subjects should be able to identify the definition of a gas and state the basic principles of Charles and Boyle's gas laws such as: Gases are forms of matter that consist of a defined number and type of particle in constant motion. While the mass of a gas is fixed, the volume of a gas is dependent on environmental factors including temperature and pressure. Gases will expand under increasing temperatures due to increased particle speed and, in turn, particle energy, and will contract under decreasing temperatures due to decreased particle speed and particle energy. Gases will also expand in low pressure environments due to decreased outside force (i.e. gravity) allowing for increased particle travel space and will condense in high pressure environments because of increasing outside force constricting particle travel space. As such, gas volume has a directly

proportional relationship with temperature and an indirectly proportional relationship with pressure.

All subjects began Phase 1 with the 12-question pretest; the measure contained all multiple choice questions. The questions were selected from the K-12 Next Generation Science standards, which delineate learning targets and exemplary questions by level of understanding. The assessment was divided into three distinct categories of rigor – low, medium, and high – based on the first three levels of the Depth of Knowledge chart: Recall, Application, and Strategic Thinking (Webb, 1997). The four low rigor questions asked subjects to identify basic concepts regarding gases under normal behaviors such as w*hat is the definition of a gas*? Questions at the medium rigor section required subjects to determine how gas behavior might change under varying conditions (i.e. *if the temperature of a gas decreases, what happens to the amount of space the gas takes up?*). The high rigor section then asked subjects to relate visible behavior changes to particle behavior at an abstract level (i.e. *how does gas particle behavior change with a decrease in temperature?*).

After the pretest, subjects were then randomly assigned into one of three learning task conditions (blended learning, guided-inquiry learning, or QBL). All subjects were then presented with the following narrative story problem to introduce them to the concept of gases and gas behavior under varying environmental conditions:

> *Suppose you decide to hike to the top of a very tall mountain and pack food for a picnic at the top. When you get there, you notice your bag of chips burst open and there are now chips all over your backpack! You think back to how the same thing happened just a few weeks ago when you*

left a bag of chips in the car on a hot summer day. You decide to investigate why this keeps happening.

After the narrative set-up, the actual learning task procedure varied by condition.

Subjects in the *blended learning condition* were directly taught the scientific principle and information through an instructional video with an accompanying video demonstration. The video was 5 minutes long and used blended instruction to convey the essential learning targets associated with the scientific principle. The first three minutes of the video are an instructor teaching a lesson via slideshow while the last 2 minutes of the video are real-world demonstrations of the phenomena. Subjects are also asked 8 follow-up questions as an attention check and to ensure the video was effective at disseminating the desired information. The distinctive feature of this condition is that subjects were told the desired learning target by an instructor and were then shown evidence of the given scientific principle occurring in a real-world environment.

Subjects in the *guided-inquiry condition* were first presented with a video demonstration of a real-world phenomenon exhibiting the larger scientific principle regarding kinetic gas behavior. They are then asked leading questions designed to direct their thinking toward a specific hypothesis. So, for example, one such video exemplifies gas behavior when temperature is changed – the following questions asked subjects to identify what happened to the size of the balloon when temperature was increased and predict why such behavior occurred. They will then repeat a similar procedure for a different video demonstrations of gas behaviors within varying environments.

As subjects move through the guided-inquiry learning task, the question focus narrows and becomes more specific until the subjects have all necessary information to anticipate the correct final hypotheses for kinetic gas laws tested in the final assessment. In all, subjects will watch three 30-60 second video demonstrations of the scientific phenomena (i.e. gas laws) and be asked 2-4 questions after each to guide the learner through the hypothesis space. Importantly, unlike in the blended learning condition, subjects in this condition are never told the desired learning target; instead, subjects are asked to generate the given learning target by predicting how and why gases behave in various ways under certain conditions.

Subjects in the *QBL condition* started their learning task with a main question screen, which included four potential questions for investigation. Subjects were told there was no specific correct answer, but to pick the question they were most curious to know in relation to the narrative story problem. The four main questions provided were as follows:

> *How do gases behave under normal conditions?*
> *How do gases behave when their environment is changed?*
> *How do we know gas particles exist if we cannot see them?*
> *Does the size and shape of the container affect how much stuff a gas is made of?*

These questions were generated using common subject misconceptions and beliefs regarding gas particles and gas behavior in past research (Jauhariyah et al., 2018; Nakhleh, 1992). Participants were then directed to a video demonstration of the selected question – a timing feature was used to encourage watching the video.

After the video, subjects were presented with the option to either ask a follow-up question related to the first question (i.e. *does changing the size and shape of a container change the amount of gas particles inside?*) or to return to the main questions screen. If subjects opted to engage with the follow-up question, they were directed to an additional video demonstration before returning to the main screen. Once back to the main screen with the original four questions, participants were free to select an additional exploration pathway or take the posttest and end the survey.

Similar to the guided-inquiry learning condition, subjects in the QBL condition are never told the desired learning target. The two conditions differed, however, in that subjects were not shown the videos in any specific order and nor were the subjects asked to answer leading questions intended to generate a specific hypothesis. Instead, subjects in the QBL condition self-selected questions to explore to help explain the phenomena observed in the narrative story problem. Table 1 outlines the methodological breakdown and major differences for each pedagogy condition for further clarity and comparison.

Table 1

Methodological Breakdown of Pedagogical Conditions

Condition	Learning Task	Major Differences
Blended Learning	Subjects will watch a 5 minute video in which the learning target is directly taught from an instructor. The final two minutes of the video are 5 real world demonstrations of the scientific principle. Subjects will answer 8 follow-up questions to encourage engagement.	Subjects are directly told the learning target (i.e. gases expand when temperature is increased) and then see examples of said learning target. All subjects see the same videos and questions in this condition.
Guided Inquiry Learning	Subjects will watch a 30-60 second video exhibiting a real world demonstrations of the scientific principle. Subjects will then answer 2-4 questions regarding what happened in the video and why they think it occurred. This repeats for 3 cycles in total so that subjects watch 5 real world demonstration videos in total (some are combined). The demonstration videos are the same as those in the blended learning condition.	Subjects are not given the learning target. Instead, they watch the demonstration videos and answer guiding questions for a specific hypothesis. All subjects see the same videos and questions in this condition.
QBL	Subjects are provided 4 exploratory questions and told to select which would be most useful for their current level of understanding. Subjects then watch a 30-60 second real world demonstration video that aims to answer the selected question. After the video, subjects have the option of selecting a similar follow-up question, returning to investigate a different exploratory question from the main menu, or opting to end the learning task. There are a total of 8 videos the subjects could view in this condition - 5 of which are the same as the prior conditions and 3 of which are not intended to be informative for the posttest following the learning task.	Subjects are not given the learning target or questions to answer. Instead, subjects self-direct the learning task by selecting one of the given exploratory questions. The number of videos watched and questions asked vary by subject selection.

Once the learning task was complete for all conditions, the same 12 question assessment given at the pretest measure was administered to generate an immediate posttest score. This experimental pretest/posttest design allowed us to determine whether a difference in achievement scores existed as a result of engaging in the condition-based learning task.

During Phase 2, we asked the same participants to complete the same 12-question measure after a week delay to determine whether delayed recall of the information differs between conditions. All subjects completed the delayed posttest measure within 7-10 days of Phase 1, which is noted in previous active learning literature as the extended time frame in which knowledge decay occurs and differences between learning models exist (Markant et al., 2016, Markant 2018).

To keep all other variables in the conditions as similar as possible, the same demonstration videos were used in all conditions and all took about the same time for subjects to complete. Any questions that were asked used similar language. All learning materials including assessment measures, video instruction, video demonstration of real-world phenomena, and guiding questions can be accessed on Open Science Framework: https://osf.io/us5eq/?view_only=3aa4c6e4e1e7440181fdc6161b380d73

Analyses

Data Preparation

Prior to running the analyses, a few additional data preparation steps were taken. Because we were interested in subject performance at the individual question level, it was

necessary to dummy code participant answers as 0 (incorrect answer) or 1 (correct answer). Once this data was cleaned and long-formatted, we had 36 observations per participant or 12 observations per time point (pretest, immediate posttest, and delayed posttest). Observations in which subjects did not provide an answer to the question (n = 3 of 5,001) were filtered out for analysis purposes; due to the low number of missing observations, we did not conduct any pre-analyses to investigate missingness. We also formatted the model to utilize the QBL condition at pretest as the reference group for the analyses as this was the condition we were most interested in investigating against the other two control conditions. All data preparation measures were done using R-programming software (R4.0.5, R Core Team).

Data Analysis

To analyze our results, we used a Bayesian logistic mixed effects model. Using the R-programming package *brms*, we regressed question accuracy (0 = incorrect, 1 = correct) on the predictor variables of condition (blended learning, guided-inquiry, QBL), time (pretest, posttest, delayed posttest), and question rigor (low, medium, high). We were interested in the two-way interaction effects between the variables of *condition*rigor* and *condition*time* as we hypothesized subjects in the QBL condition would outperform participants in the other two conditions on medium and high rigor questions and that this increased learning would hold over time.

The modeling equation used follows:

$$Score \sim condition*rigor + condition*time + (1 + rigor + time \mid subject)$$

Here, we see the dependent variable "Score" will be predicted by two fixed effects: the two-way interaction of condition and rigor and the two-way interaction of condition and time. We also included the *(1+ rigor + time | subject)* component to account for our random effects of subject variability such that subjects will respond to the learning task differently (i.e. one subject might see a small increase in performance while another subject would see a large increase in performance after the learning task) as well as accounting for the random differences in the effects of rigor and time (i.e. subjects will respond to assessment rigor and time decay differently). Finally, we include a random intercept, *(1)*, because we anticipated subjects to score differently on the measure at random (i.e. some subjects might score high at pretest while others score low at pretest).

For our purposes at this point, we used uninformative default priors included as part of the *brms* package, however, do acknowledge that future work should include informative priors to fully take advantage of the Bayesian model approach.

Results

First, we provide a summary of our data and mean results for condition group accuracy scores - scores reflect the likelihood subjects answered any given question correctly. Subjects in all conditions perform similarly at pretest (blended learning, M = 0.52; guided-inquiry learning, M = 0.50; QBL, M = 0.50). At time posttest, subjects in the blended learning condition see credibly increased scores, M = 0.70, which is then maintained at the delayed posttest, M = 0.71. Similar findings were true for subjects in the QBL condition (posttest, M = 0.68; delayed posttest, M = 0.71). Subjects in the guided-inquiry condition also see a credible increase at posttest, M = 0.60, but question

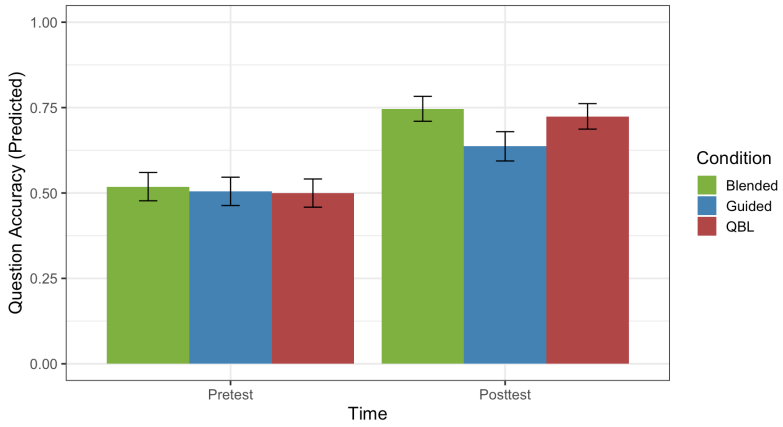
accuracy credibly differs at the delayed posttest, $M = 0.55$. We provide a detailed analysis of our hypotheses and visualized results below.

Our first hypothesis - that subjects in the QBL condition will score higher on the immediate posttest than those in the blended learning or guided-inquiry conditions - was partially supported in that subjects in the QBL condition outscored those in the guided-inquiry condition at the immediate posttest measure, $b = -.43$, $OR = .65$, 95% CI [-.83 to .05]. They did not, however, score credibly differently than those in the blended learning condition at the immediate posttest measure, $b = .03$, $OR = 1.03$, 95% CI [-.38 to .44].

We visualized our results in Figure 2. On the y-axis, we have the probability that a subject will answer a given question correctly as predicted by our model. On the x-axis, we have our time variable, which includes our measure at the pretest and posttest. The effect of condition is indicated by bar color, as depicted in the figure legend.

Figure 2

Hypothesis 1: Comparing Posttest Results by Condition



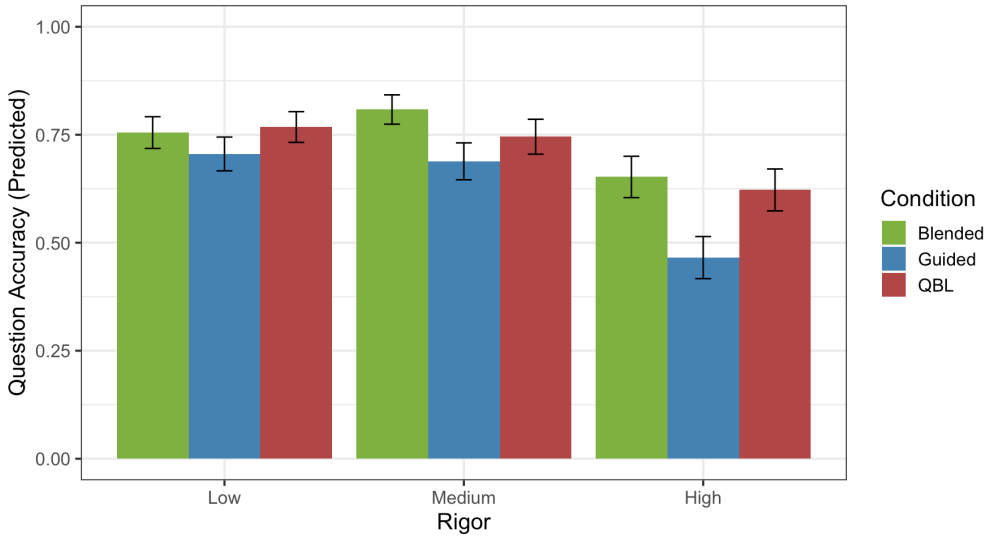
Note: A bar graph demonstrating the predicted probability a subject in each specified condition (blended, guided-inquiry, or QBL) at a given time point (pretest and posttest) will answer a question accurately. Error bars represent +/-1 standard error.

As demonstrated by the three points at time pretest, there was no credible difference between conditions in subject performance on the assessment - indicating subjects had a similar level of knowledge at the start of the survey. At time posttest, the overlapping standard error bars indicate there is no difference in predicted performance for individuals in the QBL or blended learning conditions. The probability of answering a question accurately for a subject in the guided-inquiry condition, however, is credibly below that of the QBL condition. These results provide some evidence that the QBL methodology might be more effective than current guided-inquiry models and at least as effective as existing blended learning models for learning information.

Our second hypothesis - that there will be an interaction effect for condition and assessment rigor such that subjects in the QBL model will demonstrate significantly higher scores on the medium and high rigor questions of the assessment at the posttest measure than those in the blended learning or guided-inquiry learning conditions - was also partially supported. While scores do not differ for any condition at low or medium rigor question on the posttest, subjects in our QBL condition do perform credibly better than those in the guided-inquiry condition, b = -.32, OR = .73, 95% CI [-.79 to .17] and similarly to those in the blended learning condition on high rigor questions, b = .21, OR = 1.23, 95% CI [-.28 to .70]. These results are visualized in Figure 3, with the probability our subject will answer a question accurately as predicted by the model on the y-axis, level of question rigor on the x-axis, and condition represented by bar color.

Figure 3

Hypothesis 2: Comparing Posttest Results by Rigor



Note: A bar graph demonstrating the interaction effect of rigor and condition, predicting the probability that a subject will answer a question at the given level of rigor accurately on the immediate posttest measure. Our QBL condition is represented by the red bar. Error bars represent +/-1 standard error.

We see the error bars, representing standard error bars, overlap for all conditions on low and medium level of question rigor, however, no such overlap exists for the QBL and guided-inquiry conditions on high level of rigor questions. As such, this provides some evidence for our hypothesis: the QBL model could be more effective than guided-inquiry models and at least as effective as blended learning models for teaching highly rigorous content and, therefore, increasing depth of knowledge.

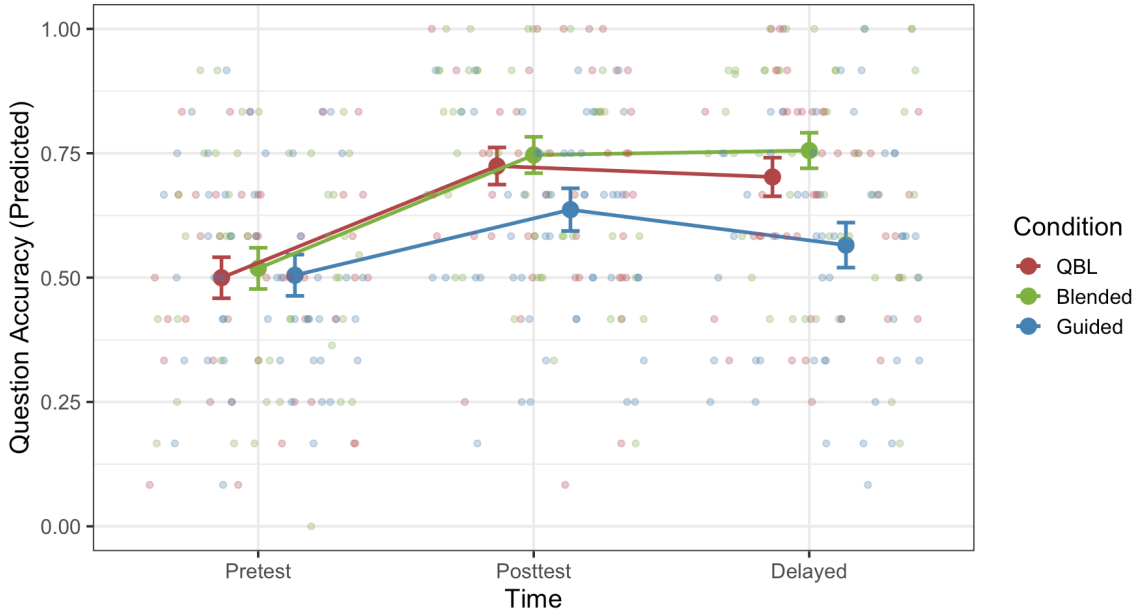
Results from Study 2 furthermore indicated partial support for our third and final hypothesis - that there will be an additional effect between condition and time such that subjects in the QBL condition will retain credibly more knowledge over time than those in the blended learning or guided-inquiry learning conditions as assessed by the week-

long delayed posttest measure. Participants for Study 2 that underwent the QBL condition learning task in Study 1 performed credibly better than those in the guided-inquiry condition on the delayed posttest, b = -.62, OR = .54, 95% CI [-1.03 to -.22]. However, performance did not differ from those in the blended learning condition, b = .19, OR = 1.21, 95% CI [-.23 to .61].

These results are visualized in Figure 4, which demonstrates subject performance over time by condition with an additional overlay to display raw mean scores for individual subjects on each measure.

Figure 4

Hypothesis 3: Comparing Knowledge Gain and Retention by Condition



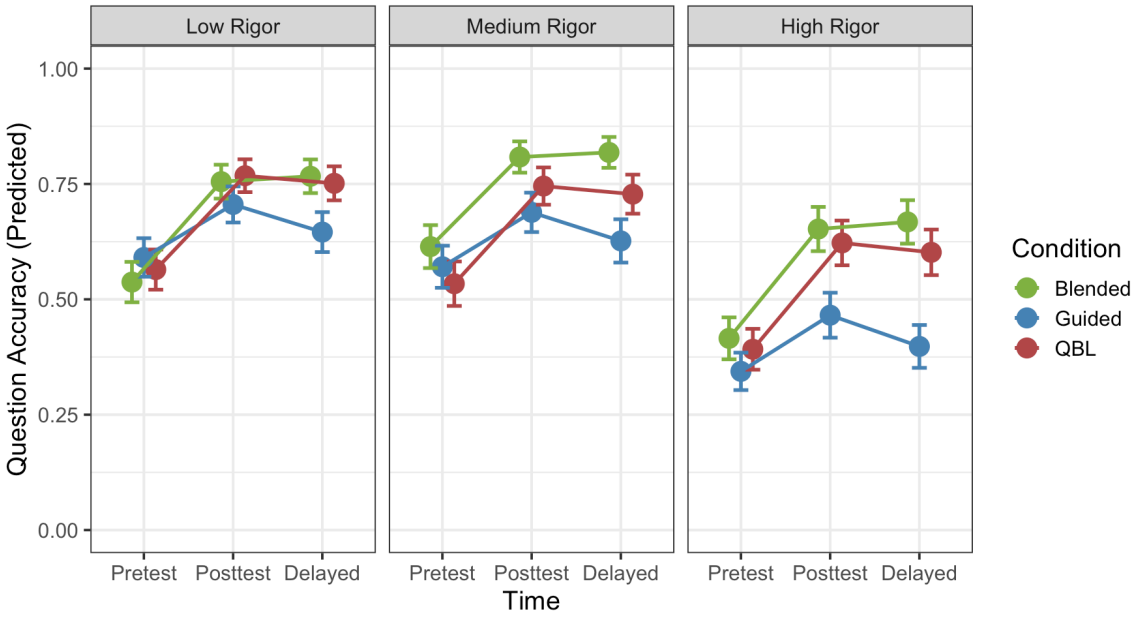
Note: A line graph demonstrating the predicted probability that a subject in each specified condition (blended, guided-inquiry, or QBL) will answer a question accurately at each time point (pretest, posttest, and delayed posttest). Additional data points represent the raw data for individual mean subject scores on the assessment at each time point. Error bars represent +/-1 standard error.

As demonstrated by non-overlapping error bars representing standard error bars at the delayed measure, subjects in the guided-inquiry (blue line) condition recalled credibly less information indicated by a lower performance on delayed posttest than those in the QBL condition (red line). The overlapping error bars for the blended learning (green line) and QBL (red line) conditions indicates that subjects in these conditions performed similarly at time posttest. Again, this provides some evidence that our QBL model is either as or more effective at reducing knowledge decay over time than existing learning paradigms.

To fully represent the entirety of our hypotheses - that subjects in the QBL model will perform better than subjects in the other two learning models credibly on higher level of rigor questions and will suffer less knowledge decay over time - we further visualized the interaction of score by condition over time faceted by question rigor in Figure 5. On the y-axis we have the probability a subject will answer a question correctly as predicted by our model and on the x-axis we have our three time points (pretest, immediate posttest, and delayed posttest). Condition is represented by separate lines, with our QBL condition in red. Finally, the graph is faceted by rigor so each graph represents a different level of question rigor (low, medium, and high).

Figure 5

Bayesian Logistic Mixed Effects Model



Note: A triptych interaction plot demonstrating the predicted probability a subject in each specified condition (blended, guided-inquiry, or QBL) at a given time point (pretest, posttest, or delayed posttest) will answer a question accurately for each discrete level of question rigor (low, medium, high). Error bars represent +/-1 standard error.

Differences in subject performance between conditions begin to emerge at the immediate posttest measure. While subjects score similarly for low and medium levels of rigor as indicated by the overlapping standard error bars, subjects in the QBL condition (red line) are more likely to answer a high rigor question correctly than those in the guided-inquiry condition (blue line) at posttest. These effects deepen at the delayed assessment. Looking at the non-overlapping standard error bars between the guided-inquiry and QBL conditions at the delayed time point, we see that subjects in the guided-inquiry condition appear to retain significantly less knowledge at all levels of rigor.

Comparing results for the QBL (red line) and blended learning conditions (green line), it appears subjects learn and retain similar levels of knowledge. Error bars at the delayed posttest for the QBL and blended learning conditions indicate the probability of a subject answering a high rigor question correctly does not significantly differ, suggesting subjects in these conditions retained similar levels of rigorous content knowledge. Interestingly, there does appear to be a significant difference in subject performance on questions at medium level rigor between the blended learning and QBL conditions. Although we did not further investigate this difference, we hypothesize this could be due to controlling access to information for subjects in the blended learning condition. All subjects in this condition received the same information and saw the same videos. The amount of information provided to subjects in the QBL condition, however, varied based on which questions were selected for exploration. We discuss the potential limitations for control of input with our QBL model versus blended learning models in our Limitations section.

Discussion

As the world rapidly modernizes, schools are pressured to produce 21st Century graduates. As such, pedagogical paradigms for science education shifted away from rote-style learning to active learning, in which students discover new knowledge through self-exploration (Chi, 2009; Chi & Wily, 2014; Levy et al., 2013). This shift in methodology increases understanding and recall of concepts – leading to improved academic performance (Markant et al., 2016, Markant 2018). But the high reward induces high

risk; active learning requires the capacity to discover, and discover correctly, the desired learning target.

Such capacity – deemed hypothesis generation – is prone to specific biases based on prior knowledge, selective memory retrieval, and failure to consider alternative explanations, and increases the likelihood students form scientific misconceptions (Bonawitz et al., 2011; Chi & Wiley, 2014; Engel, 2011; Levy et al, 2013; Markant et al., 2016; Markant, 2018). Educators, therefore, need a scaffolded intervention that reduces the cognitive load on students during hypothesis generation, but otherwise retains the benefits of active learning (Chi & Wiley, 2014; Markant, 2018, Markant et al., 2016; Preston & Eichenbaum, 2013).

To address this need, we expanded upon models of instruction for science classrooms that rely on question-asking to improve hypothesis generation. The QBL model mimics the learning methods used throughout early development: observation, experimentation, and modification (Lucca & Wilborn, 2018; Stahl & Feigeson, 2015). It also further expands upon previous work on self-regulated learning in which students are asked to recognize and effectively navigate current levels of understanding with new information (Azevedo, 2005; Azevedo, 2009). The QBL model builds this self-regulatory structure into the lesson design so that learners would be required to engage in metacognitive processing to navigate through the learning task. Additionally, our model combines work from explanation-centered learning, previously studied with computer-based tutoring, and question-asking through tutor-peer interactions into teacher-based pedagogy approaches (Grasser et. al, 2005; Roscoe, 2014; Roscoe & Chi, 2007). The

QBL model's structure, in which viable questions for autonomous exploration are provided, allows individuals to formulate, investigate, and amend beliefs as part of the learning process in a manner that is appropriately scaffolded to account for biases in hypothesis generation and current levels of individual understanding.

Our results indicate the QBL model is at least as effective as blended learning at increasing assessment scores at all levels of rigor and this learning is retained over time for low and high rigor content. Additionally, these findings provide some evidence that a QBL model is more effective than guided-inquiry at increasing performance on high level rigor questions and more knowledge gained from the learning task is maintained over time at all levels of rigor.

As such, we believe work with the QBL model should be continued and future research should focus on whether our results replicate for other scientific principles, with school-aged children, when the learning task is conducted through a non-digital platform, and against other types of commonly used pedagogical approaches. Such insight would provide additional evidence for the QBL model's effectiveness and, assuming future results support our work, validate testing the model within K-12 classrooms.

Limitations

We do recognize potential limitations in our current work. First, the QBL methodology largely assumes children would have the ability to differentiate between questions based on desired information gain. Individuals not able to do this might spend a disproportionate amount of time investigating questions that are of low information value

or potentially explore all possible questions rather than using observations to narrow their approach (Liquin & Lombrozo, 2017).

Secondly, there were many restrictions in our current work including that the model was only tested with adults rather than children, only investigated a single scientific principle, and took place on a digital platform. However, we would argue that although the current population and approach used were not ideal for our ultimate goal - designing a more effective form of pedagogy that improves hypothesis generation in K-12 science education - these same conditions were also not ideal for our tested model. It is likely our subjects are familiar with a blended learning method of receiving information, especially in a digital format; this style of technique is commonly used in various content we consume on a daily basis such as instructional videos, news stories, documentaries, etc. Using questions to explore and discover information, on the other hand, is used less frequently as a method of teaching and, therefore, likely requires a learning curve for gaining knowledge via this approach. It is possible that QBL might be more effective than blended learning if subjects are given time to practice engaging in this type of activity.

The digital platform could be additionally detrimental to the effectiveness of a QBL model; whereas the blended learning and guided inquiry models could still be structured similarly in a virtual versus classroom setting, the lack of an informed teacher might disproportionately impact the QBL approach. For both the blended learning and guided-inquiry conditions, the learning task was still largely structured and controlled because the information provided, videos watched, and questions asked were dictated by

an informed individual (the experimenter). In the QBL condition, however, this space was completely unscaffolded and dependent on the subject. It could be that subjects in the QBL condition did not see the video demonstrations needed to correctly answer questions or that these subjects selected so many questions that the video evidence was confounded over time. In a real world classroom, the teacher could facilitate the learning task in a more structured way based on independent need. We, therefore, hypothesize there could be some reason to believe the effectiveness of QBL might be even more substantial in a classroom-based environment.

An additional limitation regarding our current work is the use of multiple choice questions and a quiz type assessment as a measure of hypothesis generation. First, we cannot be sure scoring well on the given measure, such that a subject answers more questions accurately, necessarily measures hypothesis generation. In addition, while the learning targets and questions were based off of Next Generation Science Standards, using multiple choice questions could positively skew our results because it provide potential answer options for the subjects. If asked to answer in an open-ended format, we might see more misconceptions arise and less accurate hypotheses generated. Though this was a necessary limitation regarding time constraints, digital data collection, and education level of the experimenter, future work should aim to use open-ended measures that are validated and reliable for other hypothesis generation measures.

A final limitation in our work is the restricted analysis conducted with a rich data set. At the time of this writing, time restraints prohibit us from conducting further exploratory investigations regarding the nuances within conditions that are potentially

informative regarding the underlying cognitive mechanisms at work. As such, we intend to continue working with this data - specifically within the QBL condition observations - to explore how the learning task differed and how this led to increased or decreased accuracy scores.

Research questions of interest, therefore, include whether the number of questions explored and/or number of videos viewed may be positively correlated with increased accuracy scores and if whether a ceiling effect exists such that engaging with too many questions or videos might cause decreased accuracy scores. This insight would also provide some clarity as to whether subjects in the QBL condition had access to more, less, or similar amounts of information as subjects in the blended learning or guided inquiry conditions. If we were to find that subjects in the QBL condition only watched 1 or 2 videos - whereas all subjects in the blended learning and guided-inquiry learning conditions saw the same 3 videos - it would provide some plausibility that subjects in the QBL condition perform as well or better as subjects that had access to more information. On the other hand, if we find subjects in the QBL condition watch 4 or more videos on average, this might suggest that the QBL condition simply provided more information to subjects, which - in turn - increased accuracy scores. Additionally, it might be informative to explore whether asking a specific question led to higher accuracy scores; if so, this could suggest future QBL work should focus on subject's ability to ask the right questions and not just the questions they find most interesting.

Conclusion

Active learning is increasingly being utilized in K-12 science classrooms, however, pedagogical interventions are needed to improve hypothesis generation in order to maximize effectiveness. Prior work demonstrates the effectiveness of question-asking as a learning mechanism, however, there is less work regarding to to effectively implement question-asking in a systematic lesson-based structure. As such, we designed our QBL model as a pedagogical approach that utilizes student prior knowledge, hypotheses, and questions to drive learning - aiming to increase depth of content knowledge and knowledge retention and negate the formation of scientific misconceptions. Though our work is only a starting point, the demonstrated effectiveness of the QBL model as a potential pedagogical approach indicates there is reason to believe QBL could be a viable improvement to active learning practices in K-12 classrooms.

REFERENCES

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist*, *40*(4), 199-209. https://doi.org/10.1207/s15326985ep4004_2

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning*, *4*(1), 87-95. https://doi.org/10.1007/s11409-009-9035-7

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology, 45,* 180–191.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*. https://doi.org/10.1016/j.cognition.2010.10.001

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*(1), 73-105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Engel, S. (2011). Children's need to know: Curiosity in schools. *Harvard educational review*, *81*(4), 625-645. https://doi.org/10.17763/haer.81.4.h054131316473115

Fandakova, Y., Lindenberger, U., & Shing, Y. L. (2015). Episodic memory across the lifespan. In *The Wiley handbook on the cognitive neuroscience of memory* (pp.309-325). Wiley-Blackwell

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410-8415.

Fox, A. R., Hollan, J. D., & Walker, C. M. (2019). When graph comprehension is an insight problem. In *CogSci* (pp. 330-336).

Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist*, *40*(4), 225-234. https://doi.org/10.1207/s15326985ep4004_4

Jauhariyah, M. N. R., Suprapto, N., Admoko, S., Setyarsih, W., Harizah, Z., & Zulfa, I. (2018, March). The students' misconceptions profile on chapter gas kinetic theory. In *Journal of Physics: Conference Series* (Vol. 997, No. 1, p. 012031). IOP Publishing.

Jean, A., Daubert, E., Yu, Y., Shafto, P., & Bonawitz, E. (2019, January). Pedagogical questions empower exploration. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T. Y., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science*, *20*(8), 963-973.

Larsen, N., Venkadasalam, V., & Ganea, P. (2019). Without conceptual information children miss the boat: Examining the role of explanations and anomalous evidence in scientific belief revision. In *CogSci* (pp. 625-630).

Levy, B. L., Thomas, E. E., Drago, K., & Rex, L. A. (2013). Examining studies of inquiry-based learning in three fields of education: Sparking generative conversation. *Journal of Teacher Education*, *64*(5), 387-408.

Liquin, E., & Lombrozo, T. (2017). Explain, Explore, Exploit: Effects of Explanation on Information Search. In *CogSci*.

Litman, J. (2005). Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition & Emotion*, *19*(6), 793-814. https://doi.org/10.1080/02699930541000101

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin, 116*(1), 75–98.

Lucca, K. & Wilbourn, M.P. (2018). The what and the how: Information-seeking gestures facilitate learning object labels and functions. *Journal of Experimental Child Psychology, 178*, 417-436. https://doi.org/10.1016/j.jecp.2018.08.003

Markant, D. B., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education*, *10*(3), 142-152.

Markant, D. B. (2018). Effects of biased hypothesis generation on self-directed category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Nakhleh, M. B. (1992). Why some students don't learn chemistry: Chemical misconceptions. *Journal of Chemical Education*, *69*(3), 191.

Olson, S., & Riordan, D. G. (2012). Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. Report to the President. *Executive Office of the President*.

Pluck, G., & Johnson, H. L. (2011). Stimulating curiosity to enhance learning. *GESJ: Education Sciences and Psychology*, *2*(19). ISSN 1512-1801

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, *23*(17), R764-R773.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ronfard, S., Zambrana, I. M., Hermansen, T. K., & Kelemen, D. (2018). Question-asking in childhood: A review of the literature and a framework for understanding its development. *Developmental Review, 49*, 101-120.

Roscoe, R. D., & Chi, M. T. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of educational research*, *77*(4), 534-574.

Roscoe, R. D. (2014). Self-monitoring and knowledge-building in learning by teaching. *Instructional Science*, *42*(3), 327-351.

Ruggeri, A., Markant, D. B., Gureckis, T. M., Bretzke, M., & Xu, F. (2019). Memory enhancements from active control of learning emerge across development. *Cognition*, *186*, 82-94.

Ruggeri, A., Sim, Z. L., & Xu, F. (2017). "Why is Toma late to school again?" Preschoolers identify the most informative questions. *Developmental Psychology*, *53*(9), 1620.

Shah, P. E., Weeks, H. M., Richards, B., & Kaciroti, N. (2018). Early childhood curiosity and kindergarten reading and math academic achievement. *Pediatric Research*, *84*(3), 380-386.

Shing, Y. L., & Brod, G. (2016). Effects of prior knowledge on memory: Implications for education. *Mind, Brain, and Education*, *10*(3), 153-161.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91-94.

Webb, N. L. (1997). Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education. Research Monograph No. 6.

APPENDIX A

IRB EXEMPTION FOR HUMAN SUBJECT TESTING

EXEMPTION GRANTED

Nestor Pinillos
SHPRS - Philosophy Faculty
602/885-5466
pinillos@asu.edu

Dear Nestor Pinillos:

On 9/14/2017 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	The formation of intuitive scientific theories
Investigator:	Nestor Pinillos
IRB ID:	STUDY00006767
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• John Priniski Citi Quiz, Category: Other (to reflect anything not captured above); • Nestor Pinillos CV, Category: Vitaes/resumes of study team; • Zachary Horne additional citi quiz, Category: Other (to reflect anything not captured above); • Zachary Horne additional citi quiz 2, Category: Other (to reflect anything not captured above); • ASU Protocol IRB_Revision4.docx, Category: IRB Protocol; • DEBRIEFING INFORMATION SHEET.pdf, Category: Participant materials (specific directions for them); • Additional Questions and Demographic Information, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Amazon Mechanical Turk Recruitment.pdf, Category: Recruitment Materials; • Citi quiz for Samantha Roberts, Category: Other (to reflect anything not captured above);

	• Prachi CITI training, Category: Other (to reflect anything not captured above); • Science Vignettes.pdf, Category: Participant materials (specific directions for them); • PinillosCiti, Category: Other (to reflect anything not captured above); • Consent Form, Category: Consent Form; • Zachary Horne citi quiz, Category: Other (to reflect anything not captured above); • Zachary Horne cv, Category: Vitaes/resumes of study team;

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (2) Tests, surveys, interviews, or observation on 9/14/2017.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Zachary Horne
Zachary Horne
Prachi Solanki
Samantha Roberts
Nestor Pinillos
John Priniski