Running Head: INTRODUCING THE LEARNING TOOLS

Introducing the Learning Tools *Oh*

Anita E. Kelly and Scott E. Maxwell

University of Notre Dame

This article was made possible through the support from a grant from the John Templeton Foundation as part of the first author's Science of Honesty project. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. Correspondence concerning this article should be addressed to Anita E. Kelly, Department of Psychology, University of Notre Dame, Notre Dame, Indiana 46556. Phone: 574-631-7048. FAX: 574-631-8883. Electronic mail may be sent to akelly@nd.edu.

Abstract

Psychology's empirical literature has been called into question. Specific research practices in psychology have precluded the necessary evaluation of its findings. After identifying these practices, this article introduces the *learning tools Oh* for evaluating the field's findings and theories. *O* of the learning tools is a method of testing that compares two *opposite* (i.e., complementary and contradicting) predictions to see which one can be ruled out. Before a study that applies *O*, the researcher identifies a line, or fixed reference point, dividing all the relevant potential observations into those confirming one prediction or its opposite (Popper, 1959). Observations that confirm one prediction disconfirm the opposite prediction. As the universal mechanism behind a theory, *h* of the learning tools explains *how and why* one thing causes another. The presence of *h* (i.e., in a relevant context) is expected always to cause the predicted outcome. This universality of *h* allows the researcher to evaluate the theory to which *h* belongs through empirical testing. When designing a study to test a theory, *h* allows identifying a pair of opposite predictions (*O*) that would put its theory at risk. A study can falsify a theory by invoking *h* (e.g., reduced serotonin) and then obtaining the opposite finding (i.e., reduced or no change in depression) from the one expected (i.e., increased depression). By granting psychological theories the capacity to be falsified in this manner, *Oh* promises to bolster psychology's advancement of scientific knowledge.

Key words: Philosophy of science, psychological science, research methods, induction, deductive methods of testing, statistics.

Introducing the Learning Tools *Oh*

The validity of psychology's empirical literature has been called into question (e.g., Ioannidis, 2012; Makel, Plucker, & Hegarty, 2012), with its credibility said to have taken in 2012 "a quick plunge from bad to worse" (Pashler & Wagenmakers, 2012, p. 528). That same year marked the start of the Reproducibility Project (Open Science Collaboration, 2012). Psychology researchers across multiple sites attempted to replicate 100 findings published in premier psychological journals. After aggregating the results from these new studies, the researchers concluded that the overall replication rate was low (Open Science Collaboration, 2015). Thus, rather than quelling the growing concerns about psychological research, the Reproducibility Project merely punctuated the need for psychology to improve its research practices.

Since 2012, several new methodological restrictions have been proposed or mandated for research in psychology. For instance, the journal *Psychological Science* requires authors to use the *new statistics*. These statistics emphasize the reporting of effect sizes, use of meta-analysis, and replication of findings. The goal behind the new statistics is for psychology as a science to "progress toward becoming a quantitative cumulative discipline" (see Instructions to Authors, *Psychological Science*). This goal is different from the traditional role of scientists to test explanations or theories (e.g., Haeffel et al., 2009; Meehl, 1978).

Efforts to explain why psychology has produced questionable findings have put the blame, at least in part, on *confirmation bias* (e.g., Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievi, 2012). This bias refers to putting too much weight on results that confirm one's own beliefs (see Nickerson, 1998). Simmons, Nelson, and Simonsohn (2011) stated that "the culprit" behind psychology's validity problem "is a construct we refer to as *researcher degrees of freedom*" (p. 1359). Likewise, Wagenmakers et al. (2012) warned that because of

confirmation bias "considerable care needs to be taken before researchers are allowed near their own data" (p. 634).

Whereas confirmation bias can explain *why* the findings are questionable, undisclosed multiple statistical testing for a study can explain *how* the findings got this way (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011; Wetzels et al., 2011). In comparison to avoiding the extra tests, multiple testing increases the probability of obtaining false-positive results (John et al., 2012; Simmons et al., 2011; Wetzels et al., 2011). Simmons et al. (2011) conducted multiple tests with irrelevant covariates on a dataset, showing how easy it was to obtain the positive results needed to publish in psychological journals. They concluded that, therefore, psychologists have likely been conducting the extra tests.

Another reason to assume that psychologists have been conducting the extra tests is based on a recommendation that the American Psychological Association (APA) makes on its current website in the *Guide for New Authors* (APA, 2010). In the name of efficiency, new authors are advised to see the results of their tests first and then to focus the writing of a manuscript on positive findings. This advice implies leaving some of their tests undisclosed, as can be seen in Bem's (2004, p. 186) recommendation on page 10 of the *Guide*:

> There are two possible articles you can write: (a) the article you planned to write when you designed your study or (b) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (b).

Yet not reporting all of the tests and results invalidates the conclusions from a study. It guarantees that interpreting the findings will be biased. Bias is defined here as a systematic (as opposed to chance) influence on outcomes that precludes evaluating them according to the standard that was specified for that evaluation. As implied in this definition, bias pertains to how

findings are obtained and reported, rather than to what the findings turn out to be. The APA (see the *Guide* on its website) advises new authors to "consider sources of bias and other threats to internal validity, imprecision of measures, overall number of tests or overlap among tests, effect sizes, and other weaknesses of the study." These limitations are to be included in the discussion section of an empirical manuscript.

However, when bias is introduced into a study, that bias cannot be corrected after-the-fact. Bias demands specific procedures for preventing it. In a well-known philosophical book on the logic behind science, Popper (1959) offered a way to prevent bias in an empirical study. This way is to specify a standard for evaluating the outcome of a study in advance of that evaluation and to keep this standard fixed through the reporting of the outcome (Popper, 1959). The standard serves as a reference point for evaluating whether the outcome of the study confirmed or disconfirmed the predicted outcome. This process is unbiased. It allows other researchers to evaluate the outcome according to the standard specified for that evaluation, which is how the validity of a given finding is determined. Findings that cannot be evaluated are invalid.

Note that specifying a fixed standard or reference point for a given study does not imply that this standard is universal. A conclusion is reached in the context of a given evaluation of observations. What allows a prediction to be confirmed or disconfirmed in a given study is that it is being compared to a complementary prediction. The observations will show that one or the other prediction was confirmed (Popper, 1959). Other researchers are at liberty to move that reference point for future studies or to reevaluate the first study. A future study could show that the reference point should have been in a different place, and that the data that caused a theory to be rejected actually corroborated it (Popper, 1959). Having that reference point is what allows

reevaluating the data and theory. In contrast, not specifying any reference point or identifying one after examining the data introduces bias that precludes evaluating the findings altogether.

In a nutshell, evaluation requires knowledge of all the relevant data and tests conducted on them. Not implementing procedures to prevent bias and then reporting results as though they were valid would be like leaving a window open and announcing that no flies came inside. Admitting that some flies may have come inside is no solution when compared to simply leaving the window closed. This article describes methods for keeping this metaphorical window shut for the duration of a psychological study so that its findings can be evaluated.

Purpose of This Article

Thus far, we have put forward the notion that, as well as briefly explained how and why, psychology currently falls short on the necessary evaluation of its findings. The purpose of this article is to equip psychology researchers with new methods to test their theories and evaluate their data. We call these methods the learning tools *Oh*. The article first describes how to apply these tools to a psychological study. It then explains how and why these tools promise to bolster the field's capacity to advance scientific knowledge.

O of the learning tools is defined as the comparing of two *opposite* or contradicting, complementary statements to see which one can be ruled out. In a study, the two opposite statements are a pair of complementary predictions. By negating one prediction, the study's observations necessarily confirm the other. *H* of the learning tools refers to a mechanism explaining a causal relation, or *how and why* one thing causes another. For a theory that can be evaluated, or falsified, *h* is the universal mechanism behind the causal relation that the theory aims to explain. Whenever *h* is present in a relevant context, the predicted outcome is expected always to occur. Thus, whereas *O* tests a specific pair of opposite predictions at a given point in

time, *h* allows knowing what those opposite predictions should be in any study that tests the theory to which *h* belongs.

The next three sections describe how to use *Oh*. We recommend using the tools to evaluate any new set of observations after reading these sections—and before reading the last two sections. Re-reading the whole article after using the tools would be ideal, given that some of the sentences may suddenly take on opposite meanings that reflect our intended meanings. For instance, readers might initially think that we cited Popper's (1959) book throughout to enhance the credibility of our claims. But after using *Oh*, they would see that logic demands giving credit where credit is due to avoid implying that his ideas originated with us.

By using the tools, a person can expect an increase in his or her capacity to evaluate statements and observations. The reason is that *Oh* directs attention away from confirming what one considers true already and redirects that attention to evaluating what is as yet unknown. This shift allows seeing the value of "error" in "trial and error"— particularly when testing an explanation or theory, as explained later. By trying to learn what is not yet known, a person can expect to feel less defensive when problem-solving. Instead of trying to avoid looking foolish for making mistakes (e.g., saying "everyone is biased"), he or she can expect to become more immersed in the problem solving itself, taking greater risks to find more satisfying solutions.

The article is divided into five sections to follow. The first section expounds on the definitions of *O* and logic, as well as defines the terms deduction and induction. The second section expounds on the definition of *h*, explaining the role that falsifiable theories play in science. The third section illustrates how to apply *Oh* to a psychological study. The fourth section explains how and why *Oh* could provide a solution for psychology's empirical validity

problem. Finally, the conclusion hits home that by evaluating its findings and theories, psychology can expect to bolster its advancement of scientific knowledge.

Learning Tool O

Deduction versus Induction

O uses deduction to evaluate observations or statements in an unbiased manner. The reader might be accustomed to defining deduction as the application of a general rule to a specific case and induction as the generalizing of specific cases to a broader statement. These definitions merely *describe* deduction and induction without identifying the goal behind each process. In order to allow the reader to evaluate what *Oh* offers psychology, we must *explain* how and why these complementary processes are used.

Deduction refers here to finding a solution or answering a question by ruling out information (i.e., from a defined initial set of information) that would negate the correct answer (see Figure 1). This ruling out is done through one or more dichotomous comparisons of two *opposite* statements or predictions (see Popper, 1959). *Induction*, in contrast, is defined here as the gathering of positive information to confirm a given answer. With induction, observations are interpreted in a biased manner as they are compiled to justify a claim (Popper, 1959). Whereas induction obscures evaluation by aggregating observations to support a given claim, deduction allows evaluating observations by comparing them to potentially negating observations.

Popper (1959) recommended always using *deductive methods of testing* instead of induction in empirical research. In a study, *O* is the same thing as these deductive methods. A prediction is evaluated by testing whether an opposite prediction can be ruled out. *O* avoids bias by specifying in advance a fixed reference point for evaluating a study's observations. An example of *O* is the statistical significance test. It uses a sample of observations to determine

(with 95% confidence) whether the variables of interest are positively versus negatively related in a population. The reference point is the null hypothesis, or a population effect size set at zero. Finding a negative relation allows ruling out a positive relation, and vice versa.

Logic

Whichever side of the reference point the observations end up, logic allows ruling out the other side. *Logic* refers here to reasoning through deduction—not to be confused with "armchair logic" that relies on induction to make unfounded claims. Interpreting empirical findings requires deductive reasoning, given that observations cannot speak for themselves (Popper, 1963). The validity of an interpretation is established through the ruling out of competing interpretations. *O* uses logic to reach a definitive conclusion in the context of that evaluation.

Psychology as a discipline has historically pitted logic against observations, favoring observations over logic. For instance, one psychology research methods textbook explicitly ranked observations first in value as a form of knowledge (Pelham & Blanton, 2012). Logic was ranked second, followed by authority and intuition.

For this article, we borrowed this ranking forms of knowledge but made several changes. Observations were put in a separate, unranked category because they must be interpreted by another form of knowledge. Intuition and authority were combined into one "authority" category. This category is comprised of knowledge based on expertise or personal qualifications, which cannot be evaluated. The "logic" category refers to knowledge derived from deductive reasoning, where conclusions are reached through the comparing of two opposite statements. This process can be evaluated; which is why we ranked logic first and authority second (i.e., last).

The following real-life example illustrates how authority obscures evaluation, whereas logic allows it. A customer dropped the five-dollar bill that she needed to purchase a five-dollar

prescription at a CVS store. The cashier told her that the manager had just found one. But when asked for the money, the manager used her own authority to justify not giving it back. She said, "People lose money in here all the time. You said you lost it in the front of the store, and I found it in the back." Yet logic allows evaluating a decision by breaking it down to two complementary opposites specific to the context at hand. In this case, the manager could either keep the money or give it to the customer. Both parties can rule out that it was the manager's money. And if the manager was indeed keeping it for the customer who lost it, who would ever make a stronger case than the customer who just lost a five-dollar bill? The manager gave back the money.

Reference Points

Psychology researchers might be reluctant to specify their own reference points before evaluating the observations in their studies, viewing them as arbitrary. However, once they identify the reference point for a given study, *it is no longer arbitrary to those who use logic* to evaluate the findings. Once that reference point is fixed, all relevant observations can be evaluated in a nonbiased manner in relation to it. In contrast, if the researcher does not specify a fixed reference point before examining the data, that examination is necessarily biased. The researcher still uses a reference point but does not say where it is, leaving him or her at liberty to move the reference point to justify a given conclusion. Such a conclusion cannot be evaluated.

In a study using *O*, specifying the reference point for evaluating the observations would be analogous to telling a busload of concert goers to meet at Gate 2 of the concert hall by 10:15 pm to get a ride home. Once the reference point is specified, it is no longer arbitrary for that evaluation. It serves to divide the concert goers into those who are versus are not at Gate 2 by 10:15, or those who will versus will not get a bus ride home.

Another analogy is that *O*'s reference point is like the sideline, which acts as a foul line, around a tennis court. Even though scientists might argue about where to put the line, this reference point allows evaluating the shots. Anywhere in the world the tennis shots land can be evaluated with respect to the sideline. By defining which tennis shots are in versus out, the line also defines the initial set of all relevant observations by limiting them to tennis shots.

Judges sometimes make mistakes about where a ball landed. However, without a sideline, there would be no reference point for evaluating which shots were in or out, or what game was being played. The judges would have to justify any decision about who won and would be expected to have disagreements. Indeed, reviews for prestigious journals in psychology have shown poor agreement among referees (see Schwartz & Zamboanga, 2009; Suls & Martin, 2009). For example, across 153 submissions to APA journals, the publication recommendations had an intraclass correlation of only .20 (Fiske & Fogg, 1990). More essential to our point about the need for rules is that the reviewers remarked on different topics for the same manuscript.

Conclusion

O is applied to an empirical study to evaluate the observations. A fixed reference point is specified before that evaluation. It defines the initial set of all relevant observations, dividing them into observations that will confirm one prediction versus its opposite. This reference point avoids the bias that is created by seeing the results before deciding what to report. Logic allows confirming one of the two predictions by ruling out the other. Rather than relying on authority to build a case, *O* uses deduction to get to the bottom of a problem or find the answer to a question.

At first glance, this dependence on a specific context might make it seem as though the finding does not have implications beyond the sample at hand. Yet comparing observations to *O*'s pre-fixed, specific reference point allows anyone using logic to evaluate the observations in

the same way and thus to reach the same conclusion. Specifying the context for interpreting a finding would allow scientists around the world to evaluate its validity. They would understand the finding because they would know what it was compared to.

Another way that a single finding can have broad implications is when a falsifiable theory is tested and the finding confirms the opposite prediction. That one finding could cause the theory to be rejected. Although historically psychological theories have not been considered falsifiable (e.g., Shadish et al., 2001), we contend that the field could have falsifiable theories that would bolster its advancement of scientific knowledge. How and why this advancement would occur is explained in the next section.

Learning Tool "h"

Theories Exist before Empirical Testing

The last section described how to apply *O* to a psychological study. But how does one know what the reference point should be for a given study and where to put it? Popper (1959) argued that there is always a theory or empirical statement behind a study. If not, the data collection would be haphazard (Popper, 1963). As the universal mechanism behind a falsifiable theory, *h* allows defining the reference point for *O* and knowing where to put it in any study that tests the theory. This section defines theories, expounds on the definition of *h*, and describes how *h* grants a psychological theory the capacity to be falsified.

Shadish et al. (2001) observed that "once a novel and important causal relationship is discovered, the bulk of basic scientific effort turns toward explaining why and how it happens" (p. 10). Theories provide the answers to how and why, with only falsifiable theories having the capacity to be evaluated through empirical testing (Popper, 1959). Popper (1959) defined

theories as sets of statements that aim to explain and master the world, noting that "the work of the scientist consists in putting forward and testing theories" (p. 7).

Theories offering complete causal explanations have universal, falsifiable statements (Popper, 1959). An example is, "whenever a thread is loaded with a weight exceeding that which characterizes the tensile strength of the thread, then it will break" (p. 60). Theories also allow predicting what will happen in a particular case (Popper, 1959). For example, "this thread with a tensile strength of 4 pounds will break when loaded with 5 pounds." To test the theory, scientists would seek to observe a specific case that disconfirms it. A thread would be loaded with a weight exceeding its tensile strength to see if it does not break and thus falsifies the theory.

Unlike specific predictions, which can be either confirmed or disconfirmed, a theory can only be disconfirmed (Popper, 1959). Whereas outcomes from tests of predictions are simply what happened in in the past, theories identify a process expected to account for all future relevant observations. Any confirmation of a theory would fail to take into account the indefinite future tests that could still disconfirm the theory. A theory’s capacity to predict and explain future observations both makes it essential to science and precludes it from being confirmed.

Universality of H

It has been suggested that psychology cannot have falsifiable theories because most of its causal relations have multiple causes (Shadish et al., 2001). But multiple causes do not preclude a given cause from being falsifiable. *H* is a mechanism that explains *how and why* one thing causes another. *H* is considered universal when, rather than explaining every occurrence of a given outcome, its presence in a relevant context is expected always to cause that outcome.

The universality of *h* grants the psychological theory to which it belongs the capacity to be falsified by an empirical study. Instead of starting a study with a problem, such as depression,

and asking what causes it, researchers would invoke a universal mechanism (e.g., reduced serotonin) in order to test the theory. For instance, a researcher would reduce the participant's serotonin level and then assess whether that person's depression increased as expected, or showed no change or reduced depression (*O*).

If a universal mechanism is invoked and a finding confirms the opposite prediction, replacing the theory might be the obvious choice over more tests to save it. Popper (1959) wrote that more is learned from one verifiable result that contradicts a theory than from all the results corroborating it. That one result could be sufficient to reject it. Popper made a distinction between the terms *falsifiable* and *falsified*. He stated that a scientist could potentially formulate a falsifiable theory that survives rigorous tests that put it at risk, and thus it avoids being falsified.

Restricting the universality of a mechanism to when it is invoked or present addresses a complaint in the literature about studies of universal processes. Such studies have been said to assume that "people anywhere can be taken to represent people everywhere and that the cultural context of their lives can be safely ignored" (Arnett, 2008, p. 610). Yet as defined here, a universal mechanism comes with no expectation it will be invoked in every context or person. Samples are selected with the expectation that the mechanism can be invoked in at least some of the participants. The aim is to compare what happens when the mechanism is versus is not invoked, or to invoke it in all the participants to see if there are any unexpected reactions.

The testing of falsifiable theories in order to replace them with better theories offers advantages over trying to "progress toward becoming a quantitative cumulative discipline," the goal behind the *new statistics* (mentioned earlier). This goal of accumulating findings implies a preference for descriptive studies over explanatory studies. *Descriptive studies* test predictions about psychological features of a group, relations between variables, and efficacy of treatments.

Explanatory studies test mechanisms. As compared to merely describing related events, identifying a mechanism that explains how and why those events are related allows predicting when those events will reoccur (Kazdin, 2009). Moreover, descriptions of the psychological characteristics or behavior of a group do not necessarily apply to specific members of that group.

In contrast, a psychological theory with a universal *h* does allow predicting which individuals are affected by that mechanism. The universality of *h* allows testing the theory using "trial-and-error", a process normally associated with engineering. In engineering, this process ends successfully with a *trial* that solves the problem at hand. With tests of a scientific theory, however, it ends successfully with an *error* showing that the theory is wrong. Success in both cases lends insights into explanations that could account for observations that the previous theory failed to explain.

The universal mechanism, *h*, behind a psychological theory could be described with falsifiable statements that could then be tested in an explanatory study. For instance, "Necessarily, when deductive reasoning or mechanism (*h*) replaces inductive reasoning in person (*P), P* will be able to solve problems better." To test the theory, an explanatory study would invoke its universal mechanism (*h*) and then test a pair of opposite predictions (*O*). For this example, the pair of opposite predictions would start with "Person P replaces inductive with deductive reasoning (*h*)..." It would end with either "…and yet P performs worse or the same" or "…and as expected, P performed better." If the finding corroborates the theory, then it offers little information, and new studies would be conceived to put the theory at risk again.

Summary

Falsifiable theories are essential to science because only this kind of theory can be evaluated in empirical studies. A concern in the psychological literature is that multiple causes

for a given outcome, such as depression, prevent psychological theories from being falsifiable. However, as explained in this section, psychology can have such theories. To test a falsifiable theory, a study would invoke the theory's universal mechanism (*h*). *h* allows specifying *O*, or knowing where to divide the observations into those that would falsify or corroborate the theory.

Applying Oh to a Psychological Study

Explanatory Study Using Oh

When applying *Oh* to a study, the investigator purposefully puts a prediction at risk by pitting it against an opposite prediction that poses a threat to that prediction. The reason is that more is learned from the opposite finding (Popper, 1959). What follows is an illustration of applying *Oh*, versus conventional research methods in psychology, to a hypothetical study that tests a mechanism. This mechanism is that justifying, as compared to evaluating, a claim about a person reduces empathy toward that person. In the study, participants watch a video of a woman discussing moving back home. They are told, "She does not want to move back home." Then, chosen at random, half the participants are instructed to justify that this claim is true by listing 10 of her utterances that support it. The other half of the participants are instructed to evaluate whether the claim is true by listing 5 utterances that support it and 5 that contradict it. They watch the video again, list their responses, and complete a measure of empathy for the woman.

Conventional research methods. The purpose behind conventional hypothesis testing in psychology is to obtain findings that support a given hypothesis or claim. The hypothesis in this case is that "the justify group, as compared to the evaluate group, will have a significantly lower mean empathy score." After data collection, a significance test is conducted to compare this hypothesis to the null hypothesis. If the null hypothesis cannot be rejected, the researcher has the option of looking for deviant responses to drop in order to obtain a significant result. In this

example, four participants are dropped. The significance test without those four shows that the null hypothesis now can be rejected. This positive finding is reported along with these deletions, allowing the researcher to fulfill the obligation to report any dropping of participants.

However, any after-the-fact deletions hinder evaluating the findings regardless of whether those deletions were reported. For instance, they call into question whether the two groups now being compared started out equivalent to each other. They also call into question whether there were any other after-the-fact handlings of the data, such as multiple undisclosed tests that inflated the likelihood of a false positive result (see Simmons et al., 2011).

Oh. When applying *Oh* to this example, a deductive test is conducted on each of two pairs of opposite predictions. The first of these tests is the manipulation check. We chose to make the test check whether (a) 92% or more of participants in each group (Prediction M1) versus (b) fewer than 92% of participants in each group (Prediction M2) had 10 responses that matched their respective instructions. This 92% threshold for the number of participants who followed the instructions was chosen to match the level for earning an A in a college course. All participants would be retained if this threshold is met. Researchers would specify their own standards for their manipulation checks. Note that simply having a fixed reference point, regardless of whether it is selected arbitrarily, preserves the integrity of the analysis.

The second pair of opposite predictions is evaluated with a significance test. The pair is that, as compared to the evaluate group, the justify group will have (a) significantly less empathy or a non-significantly different level of empathy (expected outcome A) versus (b) significantly more empathy (opposite outcome B). Note that the null finding has to be included with one of the two predictions to make the pair complementary. It was put with A, because putting it with B would have prevented B from being the opposite outcome to be tested as intended.

Descriptive Study Using Oh

These studies assess whether certain psychological (a) traits, attitudes, or behaviors occur, (b) variables are related, or (c) innovations are effective. With a random sample from a target population, the researcher can estimate the size of an effect in that population. That estimate is reported as a specific range of values, or a 95% confidence interval when the alpha level for the test is .05. Statistical significance testing is normally considered inductive in that the tests are used to generalize results from a sample to a population. However, when the assumption behind the test of random sampling is met, deduction may be used to interpret the results. Unlike with arbitrary samples that do not refer to any population, the researcher can assume that 95% of confidence intervals will contain the population effect size.

In a hypothetical study, a random sample of 1000 Americans was asked whether they agreed with the statement, "I do not believe that 'God' exists." Based on earlier studies, the researcher used a reference point of 10% to divide the observations between her two opposite predictions. (Note that she had to pick some reference point to evaluate her data—her purpose was not to justify the correctness of this choice.) Her predictions were that "more than or the same as the expected 10% of this sample will agree" (Prediction A) versus "fewer than the expected 10% of the sample will agree" (Prediction B). The statistical significance test showed that Prediction B was confirmed, where 8% of the sample agreed, which was statistically significantly different from 10%. Thus, the researcher reported that "fewer participants than expected said that they do not believe in God."

The sufficiently large random sample from the population allowed the researcher to use logic to rule out having less than 95% confidence that the finding applied to the population. Thus, she concluded that "I am 95% confident that if asked, fewer Americans than the expected

10% would say that they do not believe in God." Her conclusion was valid in the context of that evaluation. If it is later discovered that the reference point should have been 6% instead of 10%, the capacity to evaluate that same set of observations remains intact. The reference point can be moved to 6% to make the new conclusion valid in the new context. Upon seeing that 8% is statistically significantly greater than 6%, the researcher concludes that "I am 95% confident that if asked, more Americans than the expected 6% would say that they do not believe in God."

How Oh Would Change a Psychological Study

Applying *Oh* to a psychology study would change how the literature review is written, how a study's predictions are stated, what sampling procedures are used, how the data are analyzed, and how the findings are interpreted and reported. First, the literature review would not justify the study, but rather would make clear what was being studied by defining all key terms and explaining the logic behind the design and tests. Second, pairs of opposite predictions (*O*) would be stated in advance, with a fixed point of reference dividing potential outcomes confirming one prediction or its opposite (Popper, 1959). When a mechanism (*h*) is tested, manipulation checks of whether it was invoked would be included in the predictions.

Third, the sample selected for the study would not violate assumptions of the statistical tests. If the sample is to be used to make claims about a population, then random sampling from that population is required. Fourth, the findings would be evaluated and reported in the context of the two opposite predictions. For example, "The results confirmed Prediction A, thus disconfirming Prediction B." The analyses would correspond to tests of the predictions. No undisclosed multiple tests would be permitted. All the data would be retained for the analyses, and all results would be reported. When a prediction is confirmed that contradicts the prediction from the theory, efforts would be made to verify the findings rather than to defend the theory.

Comparing Induction and Deduction for Psychological Studies

Thus far, the article has explained how and why the learning tools allow evaluating findings and theories, as well as how to apply them to a psychological study. Therefore, readers have the information needed to use *Oh* to evaluate of a new set of observations. We ask readers to stop here to apply the tools. As mentioned earlier, *Oh* can cause readers to interpret many of the sentences on these pages in new ways that line up with the meanings we intended.

Transformation from Induction to Deduction

Readers who try *Oh* may feel a transformation as they shift from inductive to deductive thinking. With induction, observations are gathered to support a claim. The pile gathered is then compared to no observations, which can convince a person that the claim is true. Metaphorically speaking, the process starts from the ground and works its way up. Induction obscures evaluation by focusing attention on the accumulated evidence, which is comprised of irrelevant and/or a subset of the relevant information. Accumulating evidence lines up with the notion that "there exists no single physical, social, or psychological reality. Multiple realities coexist, and it is up to individuals to learn about and accept these multiple realities" (Kernis, 2003, p. 15).

In contrast, deduction is consistent with the notion that only one reality exists. It is up to individuals to specify their theories and put them at risk by testing predictions from those theories within specific contexts. Deduction starts at the top, identifying the entire set of relevant observations, and works its way down. Logic allows getting to the bottom of a problem. Switching from induction to *Oh* shifts attention from some to all of the observations that must be evaluated to reach a definitive conclusion in a given context. Rather than describing a member of a category with terms that are descriptive of that category, each member would be evaluated on a case-by-case basis. The relevant mechanism explaining how and why (*h*) one thing causes

another would be assessed, along with whether or not (*O*) that mechanism is operating in that case. By calling attention to negating information that would be missed using induction, *Oh* allows understanding how and why justifying claims is biased.

Specifying versus Not Specifying the Reference Point

Specifying a fixed reference point between observations that would confirm one prediction or the opposite is what allows deduction to start from the top to reach a definitive conclusion. It does so by defining the initial set of all relevant potential observations. The observations are then compared to the reference point to rule out the wrong prediction.

Without specifying the reference point, a quantified finding from a study cannot be evaluated. An example is saying “most people” without providing a context for interpreting “most”. Another example is “most of our sample” when only a subset of the data was reported.

One of the new mandates in psychology is to report effect sizes for any given study (Cumming, 2013). A seeming benefit is that the effect sizes of findings from different research teams testing similar questions can be aggregated, with the results from individual studies combined in a meta-analysis. An example of such a meta-analysis is the Registered Replication Report (Alogna et al., 2014). Meta-analysis is described in favorable terms by the new statistics. They state, “We are most likely to be convinced an effect is non-zero after the initial study has been replicated, and a meta-analysis of all relevant studies gives an overall estimated effect size, with confidence interval, that we can interpret as indicating a non-zero effect.”

Yet when individual findings cannot be evaluated, combining them in a meta-analysis does not make them valid. Meta-analysis increases confidence in a given effect size. However, it does not allow evaluating whether that effect size approximates the true or population effect size.

Stating in advance a fixed reference point would allow evaluating the observations in a study. It is a simple procedure. And yet it could cause a shift in the researcher from using induction to deduction, which would be a dramatic change in orientation from piling up evidence to evaluating it. This change would lead to better psychological studies, with "better" defined as going from "cannot be evaluated" and "relies on authority" to "can be evaluated by using logic."

When relying on authority, as opposed to logic, researchers use induction to justify claims. Trying to generalize results beyond their samples compels them to obscure evaluation of their methods, tests, and findings. For instance, psychological studies with convenience samples yield findings that pertain only to the samples at hand. If the researchers were to lay out the logic behind the selection of participants and the assumptions of the significance test, it would become evident that no claims could be made about a population.

Switching to logic can be expected to change how researchers view the literature review prior to a study, every research decision for the study, and how they interpret their findings. Instead of trying to round up evidence to support predictions from their theories, they would evaluate their predictions through pitting them against competing predictions. They would draw conclusions in the context of a given evaluation, rather than generalizing beyond what they have evaluated. Instead of viewing the observations from samples as not worth much, they would devise ways of making the observations from their samples tell a bigger story through tests of theories. Falsifying accepted theories would allow replacing them with better theories.

The new statistics aim for psychology to become a cumulative discipline of descriptive, rather than explanatory, findings. Yet without testing theories, psychological studies can only yield results that are snapshots of the past. Accumulating these snapshots would be like piling up

drawings of unrelated cartoon characters. In contrast, theories with a universal mechanism would connect the metaphorical cartoon drawings, bringing the characters to life.

Conclusion

Specific research practices in psychology preclude evaluating the field's findings. This article introduced the learning tools *Oh* as methods for evaluating psychological findings and theories. These tools use logic, defined as deductive reasoning, to obtain a valid finding within the context of evaluating two opposite predictions (*O*). By using *O* to test against the opposite of what is expected, a study can extend knowledge into what is currently unknown.

H refers to a process or mechanism that explains how and why one thing causes another. When a theory has a universal *h*, invoking it allows specifying a pair of opposite predictions (*O*) to test the theory in a given study. *H*'s universality is defined by the expectation that, rather than always causing a given outcome, *h* would cause that outcome whenever *h* was present in a relevant context. This new conceptualization of the universality of a mechanism would allow psychological theories to be falsifiable, or evaluated by empirical tests. By obtaining a finding that disconfirms a prediction from a falsifiable theory, scientists can glean insights into a better theory to replace that theory (Popper, 1963).

The learning tools *Oh* offer a simple solution for psychology's empirical validity problem. Using *Oh* requires making explicit the role of logic in research, which has not been explicit. Psychologists have had to rely on authority instead of logic to interpret findings. Yet, like the CVS manager's justifications for keeping the five-dollar bill, the use of authority cannot be relied on to be fair. Using *O* allows seeing that every logic-based decision can be boiled down to choosing between two complementary alternatives. Choosing a given alternative does not require 100% certainty. It only requires being more certain of that alternative than of the

complementary one. *Oh* also makes explicit the role that theories play in research. Historically, scientists across disciplines have downplayed their theories until they could gather enough evidence to support them (see Popper, 1963). With *Oh*, however, they would specify their theories up front for the purpose of evaluating them and drop those theories whenever falsified.

Addressing Limitations

Psychology has introduced the "new statistics" to address its empirical validity problem These statistics emphasize the accumulation of descriptive findings, implying a de-emphasis on the testing of theories. Some readers might prefer the new statistics over the testing of theories or using logic in research. Descriptive findings are concrete, after all, with predictions that can be confirmed or disconfirmed. We argue, however, that without putting forward and testing falsifiable theories, accumulating descriptive findings merely piles up snapshots from the past. Psychologists would still have to formulate theories. But their theories would not be falsifiable. Thus, psychologists could only use findings to justify, rather than to evaluate, their theories.

For readers who do test theories, some might wonder, "Why should I have the goal of disconfirming my own theory?" Our response is that the alternative would be to defend a theory, which introduces bias to its tests. Rather than thinking of a theory as belonging to a person, we think of it as a framework for scientific discoveries to be shared with interested parties. When applying *Oh*, the goal is to evaluate the theory, with clever tests devised to put it at risk. The ideal study obtains a result negating an accepted theory after invoking its universal mechanism.

Finally, readers might use *Oh* and enjoy its benefits, just as described, but find that conflicts arise with colleagues using induction. These conflicts are to be expected, given that induction and logic lead to opposite conclusions. Without logic's capacity to evaluate observations and theories, induction creates a metaphorical path to nowhere. Those colleagues

using induction might argue that the null hypothesis is never true, observations always come with measurement error, past behavior is the best predictor of future behavior, psychology cannot have falsifiable theories, and it will never be a hard science. But logic allows seeing that the null hypothesis can be true, observations do not necessarily have measurement error, and a falsifiable explanation of how and why a behavior occurs allows accurately predicting its re-occurrences. We have demonstrated that psychology can test falsifiable theories, just as the natural sciences do. Using *Oh* would allow psychology researchers to evaluate these claims.

Take-home Message

Dichotomous tests (*O*) of theories with a universal mechanism (*h*) hold the promise of bolstering psychology's capacity to advance scientific knowledge. For a given test, *O* allows interpreting the result in comparison to the opposite result that did not happen. *O* can always be used instead of induction in an empirical study. Across many tests, the universal mechanism (*h*) allows any one result to show the theory to be false.

A falsifying outcome not only implies a rejection of a theory, but also lends insights into a theory to replace it. The mechanisms behind the replacement theories can be expected to become progressively more fundamental, extending to a broader range of circumstances. One can imagine how scientific knowledge would expand if scientists across all disciplines had the objective of discovering the limits of existing theories in order to invent better theories. Given that induction is a psychological phenomenon (i.e., a mental process that has social and emotional consequences), psychology could lead the way for such a change across the sciences. Applying *Oh* would be expected to allow scientists in any discipline to shift from justifying their theories to evaluating them instead.

References

Alogna, V. K., et al. (2014). Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556-578.

American Psychological Association. (2010a). *Preparing manuscripts for publication in psychology journals: A guide for new authors.* http://www.apa.org/pubs/authors/new-author-guide.pdf. Retrieved 4/27/16.

American Psychological Association. (2010b). *Publication manual of the American Psychological Association (6th ed.).* Washington, DC: American Psychological Association.

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist, 63,* 602-614.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323-370

Bem, D. J. (2004). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The compleat academic* (2nd ed., pp. 185-219). Washington, DC: American Psychological Association.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews: Neuroscience, 14,* 365-376.. doi:10.1038/nrn3475

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science.*

Fiske, D. W., & Fogg, L. (1990). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. *American Psychologist, 45,* 591-598.

Haeffel, G. J., Thiessen, E. D., Campbell, M. W., Kaschak, M. P., & McNeil, N. M. (2009).

Theory, not cultural context, will advance American psychology. *American Psychologist, 64*, 570-571. DOI: 10.1037/a0016191.

Ioannidis, J. P. A. (2012). Why science is not necessarily self correcting. *Perspectives on Psychological Science, 7*, 645-654.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science, 23*, 524-532.

Kazdin, A. E. (2009). Understanding how and why psychotherapy leads to change. *Psychotherapy Research, 19*, 418-428.

Kernis, M. H. (2003). Toward a conceptualization of optimal self-esteem. *Psychological Inquiry, 14*, 1-26.

Makel, M., Plucker, J., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*, 537-542.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147-163.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting & Clinical Psychology, 46*, 806-834.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175-220.

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657-660.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*, 1-8.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science : A crisis of confidence? *Perspectives on Psychological Science, 7,* 528-530. DOI: 10.1177/1745691612465253

Pelham, B. W., & Blanton, H. (2012). *Conducting research in psychology: Measuring the weight of smoke.* Wadsworth, London.

Popper, K. R. (1959). *The logic of scientific discovery.* London: Hutchinson.

Popper, K. R. (1963). *Conjectures and refutations*. Routledge.

Psychological Science (2016). *Instructions to authors*.

Rapport, M. D., Bolden, J., Kofler, M. J. Sarver, D. E., Raiker, J. S., & Alderson, R. M. (2009). Hyperactivity in boys with Attention-Deficit/Hyperactivity Disorder (ADHD): A ubiquitous core symptom or manifestation of working memory deficits? *Journal of Abnormal Child Psychology, 37,* 521-534.

Schwartz, S. J., & Zamboanga, B. L. (2009). The peer-review and editorial system: Ways to fix something that might be broken. *Perspectives on Psychological Science, 4,* 54-61.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science,* 9, 59–71.

Suls, J., & Martin, R. (2009). The air we breathe: A critical look at practices and alternatives in the peer-review process. *Perspectives on Psychological Science, 4,* 40-50.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632-638. DOI: 10.1177/1745691612463078

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*, 291-298.

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4*, 49-53.

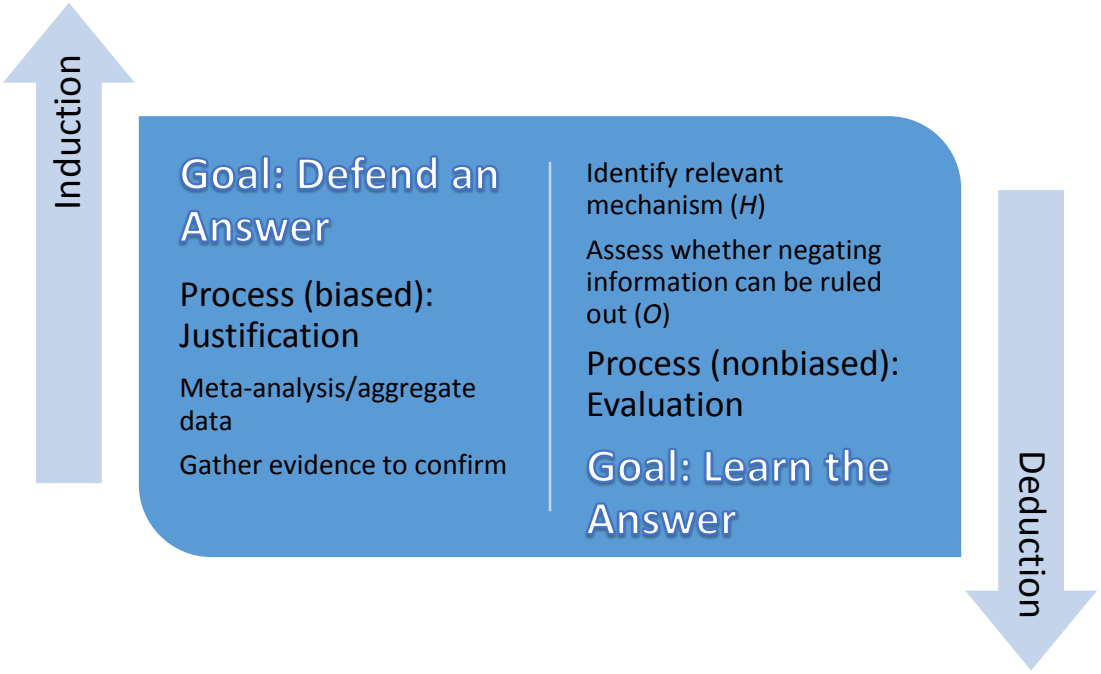

Figure 1. *Oh* in the context of using deduction versus induction.