Barron, K.E., Brown, A.R., Egan, T.E., Gesualdi, C.R., & Marchuk, K.A. (2008). Validity. In S.F. Davis & W. Buskist (Eds.) *21st century psychology: A reference handbook* (pp. 55-64). Thousand Oaks, CA: Sage.

7

VALIDITY

KENNETH BARRON, WITH ALLISON R. BROWN, THERESA E. EGAN, CHRISTOPHER R. GESUALDI, AND KIMBERLY A. MARCHUK

James Madison University

If we had to identify one concept that we hope all psychology students would master after taking their coursework in research methodology, it would be *validity*. Put simply, validity represents accuracy, credibility, and soundness. It is derived from the Latin word *validus*, which means strength. When applied to psychology, validity provides a vital tool to judge the quality of our scientific evidence about psychological phenomena.

However, despite the important role that validity plays in psychology, it can be easy for new students to get lost in the different terms and jargon associated with it. Therefore, the purpose of our chapter is twofold. We begin with a class exercise that we use to introduce validity and its role in the scientific process. Then, we review the common validity terms that appear in psychology and advance an overarching framework of validity to simplify and organize the different terms and jargon associated with validity.

AN INTRODUCTORY EXERCISE ON VALIDITY

As an initial exercise to introduce validity, we ask students to read the following article from the University of Wisconsin's school newspaper, *The Badger Herald*. So you can experience this activity firsthand, a shortened version of the article appears below. Take a few moments to read the article, and we'll pose to you the same questions we ask our students.

A Recent Survey Conducted by the Princeton Review Named UW, Madison, the Number 2 Party School in the Country*

By Kira Winter, Feature Editor

Welcome to UW, Madison—the No. 2 party school in the nation. UW was recently bestowed this high honor by the Princeton Review, finishing only behind West Virginia University, Morgantown.

UW lived up to a "study hard, party hard" reputation in the minds of 56,000 surveyed students.

Although UW Junior Lindsay Moore does not disagree that she attends a school where students like to have a good time, she is leery about the reliability of the findings. "There are surveys for just about everything," Moore said. "I think if someone came here on a football weekend they would see a very different campus than if they came in the dead of winter."

Moore's mother, Gail, thinks students at every college or university party, but UW's heritage singles it out. "Drinking, especially beer drinking, is definitely part of the Wisconsin tradition and people play into that," she said.

How do incoming freshmen feel about maintaining the school's wild reputation? "I think it's great that UW was voted the number two party school and I am definitely excited to contribute," said freshmen Andrea Santiago.

*Winter, Kara. (1997, August 29). A recent survey conducted by the Princeton Review named UW-Madison the number 2 party school in the country. *Badger Herald,* University of Wisconsin, Madison. Reprinted by permission of the *Badger Herald.*

> And what about those concerned parents, leaving their innocent children in the clutches of such wide-eyed hedonism? Dan and Joanne Schwartz were slightly apprehensive about leaving their son, an entering freshman, after they heard the survey on the news, but think UW students maintain a good balance of studying and partying. "To be honest, we had a hard time sleeping the night we heard that it was the number two party school on the national news," said Schwartz's father. "But we know it's a great school academically, and fun, too, so we really aren't too worried."
>
> Many school administrators are taking the new honor with a grain of salt. "Of course the issue of alcohol use is something we are definitely concerned with," said Roger Howard, Associate Dean of Students. "However, I don't think excessive alcohol use and being voted the number two party school are really related. It has more to do with the enthusiasm and energy of the students."
>
> Top 10 Party Schools: 1. West Virginia University, Morgantown, 2. University of Wisconsin, Madison, 3. SUNY, Albany, 4. University of Colorado, Boulder, 5. Trinity College, 6. Florida State University, 7. Emory University, 8. University of Kansas, 9. University of Vermont, 10. Louisiana State University.

Now, consider the following questions: Is the statement that UW is the No. 2 party school in the nation *valid*? If yes, which elements give the article credibility to support this conclusion? If no, which elements have you questioning the conclusion that UW is the No. 2 party school?

When we conduct this exercise, students identify several things from the article that appear to support the validity that UW is the No. 2 party school. For example, they note that the findings are based on a large sample of 56,000 students. They also note that the study was conducted by the Princeton Review, which students regard as a credible source because it is well known for publishing college guidebooks and SAT prep books. However, they also identify several things that call into question the validity of this finding. For example, they note that one of the students quoted in the article stated that a different impression of the "party" atmosphere of the Wisconsin campus could be formed depending on when you visited the campus. Students also indicate that the label of "party school" may carry different connotations for different people, thereby influencing results. For example, although the UW students and parents interviewed in the article appeared to connect a "party school" to the consumption of alcohol, the Associate Dean of Students noted that UW students are enthusiastic and energetic. We will return to this article as a case study throughout the chapter.

WAYS OF KNOWING

Interestingly, before critically discussing the article and questioning the validity of its findings as a class, many students admit that they accepted the conclusion that UW was the No. 2 party school. Charles Peirce, a 19th-century American philosopher of science, outlined several different strategies that individuals employ as they come to accept something as true. (These have been popularized as the "ways of knowing.") One way is through *tenacity*, which is the acceptance of a belief based on the idea that we have always known it to be that way. For example, when reflecting on our newspaper article about UW being the No. 2 party school, one of the parents interviewed noted, "Drinking, especially beer drinking, is definitely part of the Wisconsin tradition and people play into that." When beliefs are based on tenacity, they are difficult to change, but when individuals are objective and open to new data, these beliefs may be disconfirmed.

A second way of knowing is through *authority*, which means accepting a belief because authority figures (such as parents, teachers, or experts) tell us to. In the case of UW being No. 2, the article was based on information reported by the Princeton Review, which could be deemed an authority about university life. However, authority figures may be wrong or have hidden agendas.

A third way of knowing is through *rationalism*, which is the acceptance of a belief through the use of logic and reasoning. Again, one may logically deduce a lot of partying occurs at UW through deductive reasoning—for example, people party at bars. Madison, Wisconsin, has lots of bars. Therefore, Madison has lots of parties. Rationalism has appeal because we are now using our own independent thinking and logic to form an opinion. However, what may make sense in theory may be far from true in reality.

A fourth way of knowing is through *empiricism*, which is the acceptance of a belief based on direct personal observation. If you haven't been to UW, a visit could reveal many highly visible places in which you could observe students engaged in partying. For example, at the time that this article was published, UW's Student Union still served alcohol to students, faculty, and staff. However, direct experience also can be biased and misleading. Note how one student quoted in the article indicated you might get a far different impression if you visited on a football weekend versus another time of year, especially in the dead of winter.

However, Peirce was quick to point out how each of the previously mentioned ways of knowing can be misleading. Each can be subjective and influenced by idiosyncratic or cultural biases, they can be marked by an uncritical and accepting attitude to gaining knowledge, and, in the case of the first three ways of knowing, no systematic search for verifiable truth is made by the individual. Thus, knowledge obtained from these methods may or may not be true, but you won't have any way of telling the difference. His solution was to use *science* as a fifth and final way of knowing,

When adopting a scientific approach, we combine rationalism and empiricism. Rationalism (using logic and reasoning) is related to theory-building and generating hypotheses. Science often begins by devising a "theory" to explain a particular phenomenon. Empiricism (using personal observation) is related to testing theories and hypotheses. As a way of knowing, a scientific approach

systematically evaluates whether the prediction of a given theory will hold true in the observable world. "As one leading psychological researcher put it, the scientific method...helps us sort out what we know about human nature from what we only *think* we know" (Rosnow & Rosenthal, 1996, p. 6).

JUST BECAUSE YOU ENGAGE IN SCIENCE DOESN'T MEAN IT'S GOOD SCIENCE

However, just engaging in the scientific method as a way of knowing doesn't guarantee good science! We still need a framework to determine if a study's conclusions are valid and to appreciate that not all science is "created equally." Scientific studies may vary tremendously in the quality of information and strength of evidence that they provide. Thus, another defining feature of science is to adopt a skeptical and critical attitude to judge the quality of evidence that we collect. Concepts of validity provide that framework, and if you take the time to master those concepts in your psychological training, you will place yourself in an excellent position to strengthen your ability to separate good science from bad science. It also will provide you with the ability to determine what science is lacking and what should be done to improve on what we currently know.

Mastering validity, however, requires learning a new vocabulary. Validity is an area of science in which a wide range of terms is commonly used. For example, we conducted a review of 15 popular undergraduate psychology research methods textbooks to evaluate how many terms are typically reported. Individual textbooks described from 4 to 12 different types of validity, and 15 unique types of validity were discussed across all textbooks. Making matters worse, you may encounter researchers who use *different* terms to describe the *same* validity concepts. Or you may encounter researchers who use the *same* terms even though they are talking about *different* validity concepts.

A common critique of many psychological fields concerns the proliferation of terms used to describe various concepts, especially as work in that field grows over time. Thus, an important call for any field is to balance the number of terms and concepts used to capture the complexity of the phenomena with striving for parsimony and the simplest explanation (which in this case would be the fewest validity terms necessary). In addition, if we can avoid redundancy and overlap in our language, we can develop a simple, common language that can be quickly learned, providing a universal dialogue in which to converse.

It is with this goal that we would like to simplify the complexity found in validity terminology. Following Shadish, Cook, and Campbell's (2002) guidelines, we recommend a validity framework in which the overarching themes of validity can first be grouped into four major areas. Then, we will share our review of the research literature on validity for other terms that have appeared that can fall under the umbrella of one of these four overarching areas. With our students, we affectionately refer to this organizational framework as the Big 4 model of validity (modeled after the Big 5 model in personality). These areas entail *construct validity*, *external validity*, *internal validity*, and *statistical conclusion validity* (see Cook & Campbell, 1979; Shadish et al., 2002).

CONSTRUCT VALIDITY

Construct validity involves making a systematic evaluation of the accuracy and strength of the constructs used in the study. A *construct* is simply a synonym for the research variable being investigated. Thus, construct validity entails asking the following question about a study: "Does the study accurately represent and evaluate the constructs (aka variables) under investigation?"

To diagnose if a study is weak or strong in construct validity, you first need to determine each of the constructs that the researcher is interested in studying. Then, for each of these constructs, you want to make an independent judgment on whether you think the researcher has accurately represented that construct. For example, in our opening case study, the Princeton Review was interested in finding out how universities differed with respect to student partying. In this case, there is only one main construct being studied: partying behavior. However, an essential component to any scientific research study is converting a research question about a theoretical idea (like partying) into a measurable way to assess and evaluate the construct. When we critically evaluate the construct validity of our opening article, we need to consider how the researchers defined and measured partying.

Turning back to the article about UW, can you determine what conceptual definition the Princeton Review used for partying, and what actual measure the Princeton Review used to determine it? Unfortunately, when the media reports research, detailed descriptions of the actual measures used are rarely included. Instead, we are left to brainstorm the possible ways in which partying behavior could have been measured. For example, does partying mean the number of social events a student attends while in college? Does partying mean the number of actual hours a student spends partying? Or does partying mean engaging in a certain kind of activity, like drinking alcohol, which may be in fact what the Princeton Review wanted to capture? Depending on which definition of partying is used, we may find a rather different answer regarding which university students' engage in the most partying. So, we might quickly critique that "party school" is an ambiguous term that should be replaced with a more descriptive term such as "alcohol use" or "social drinking," if that was the goal of the Princeton Review's study of college campuses.

Unlike popular media sources, academic reporting of research involves a much more detailed and in-depth explanation of what a researcher did to make clearer evaluations. Researchers also have many additional types

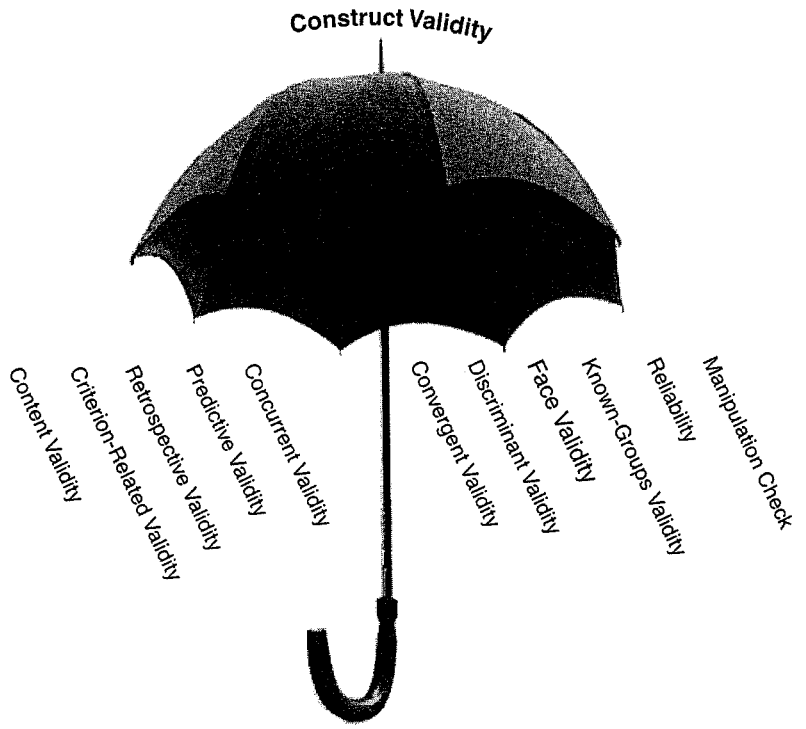

Figure 7.1 Specific terms and concepts associated under the umbrella of construct validity.

of validity and types of research tools that help them argue that they truly have accurately represented the psychological constructs that they are interested in investigating. Indeed, there is an entire psychological field focused on developing tools for the accurate measurement of psychological phenomena (Benson, 1998; Bryant, 2000; Cronbach, 1990; Messick, 1995; Shepard, 1993). To help simplify the complexity of what you may encounter and read, we would like to place these additional terms under the broader umbrella of construct validity (see Figure 7.1). These concepts are subtypes of construct validity. As researchers, we look for this type of information in academic journals when we critique the construct validity of a study. We want to see the researcher convince us that the measures that have been adopted will provide a strong and accurate measure of the constructs that the researcher is investigating.

There is no doubt that construct validity is the most complex of our Big 4 validities to describe because of its many different subtypes. So, let's take another example of a measured variable important for getting into many colleges, the SAT, which is a construct that is argued to reflect students' college preparedness and potential academic aptitude. This example will help you appreciate how researchers evaluate additional information about a measure in order to argue that it is high or low in construct validity.

One source of evidence is *content validity*, which reflects the extent to which a measure appropriately covers the full breadth and range of the construct under investigation. If you use a scholastic aptitude test to measure what you've learned in school, then what content should be on that test if you want to argue the measure is high in content validity? Should it just have math, verbal, and writing components like the SAT? Or should it reflect the broader range of academic subjects that you take in school (including content from the natural sciences, social sciences, arts and humanities, and foreign languages)? Right away we may critique the SAT for being a narrowly focused test. Thus, it lacks content validity, which has stimulated the development of other, broader-based tests such as the ACT. The narrow focus of the SAT is one reason why colleges and universities often ask students to take other achievement tests.

Another source of evidence is *convergent validity*, which is a technique that involves comparing your measure to other known measures to which it should be theoretically related. For example, the SAT could be compared to other known college entrance assessments such as the ACT, and we would predict that students' scores should be positively related if in fact both reflect the same underlying theoretical constructs. However, equally important can be demonstrating *discriminate validity* by showing how your measure is distinct and unique from other measures. For example, if we wanted to demonstrate that the SAT reflects a specific aptitude test rather than a more general IQ test of overall intelligence (another popular notion of what the SAT evaluates), we could compare students' scores on the SAT scores to IQ scores. If we are correct, SAT and IQ scores should be less related to each other than SAT and ACT scores. By showing how our measure both relates to and doesn't relate to other known measures, we are in a better position to validate what our construct represents.

Another source of evidence is *criterion-related validity*, which reflects the extent to which a measure accurately relates to other criteria or outcomes to which you would theoretically predict it would be related. These other criteria may be about things in participants' past, present, or future. Researchers will then compare scores on their measure with these criteria to see if they align with their predictions. For example, SATs could be compared to students' high school GPA to look at criteria in the past (this type of criterion validity is known as *retrospective validity* because you are going back in time retrospectively to acquire information that should be related to your construct). We might predict that if you do well on the SAT, you also probably did well in high school. Alternatively, the SAT score could be compared to students' future college GPA because you might predict that students with higher academic aptitude coming into college would perform better in college (this type of criterion validity

is known as *predictive validity* because you are going forward in time). The SAT score also could be obtained at the same time with other measures that you believe are associated to the SATs, like the ACT (this type of criterion validity is known as *concurrent validity* because measures are all collected at the same time).

Face validity evaluates how accurate a measure seems on its appearance to measure the construct of interest (its face value). So, after reviewing test items on the SAT, you would likely conclude that it appears to assess math and verbal aptitude. If the items assessed other content but the researchers still argued that it only assessed math and verbal aptitude, you might question the face validity of the test. Finally, *known-groups validity* is a technique in which a researcher gives a measure to known groups that should differ on the measure. For example, scores on the SAT should be higher for a group of students who take college prep work in high school, and scores should be lower for a group of students who dropped out of high school. If the groups differ in the predicted way, then the measure will appear to be more accurate. Although there are many different subtypes of construct validity, the take-home message is appreciating that these additional validity types are research tools to aid us in making an overall judgment about construct validity.

The accurate measurement of constructs also is tied to a related concept known as *reliability*. Although reliability often appears as its own topic in research method textbooks, we argue that it, too, falls under the broader umbrella of construct validity. Whereas validity is associated with overall accuracy and strength, reliability is associated with consistency. Researchers wish to make sure their measures consistently assess the behavior in which they are interested to assure that their measures are valid. If their measures are unreliable, then the construct validity of their measure is quickly jeopardized. Just as there are many subtypes of validity, there are many subtypes of reliability that researchers report to help support that they have collected a reliable measurement. The types of reliabilities that are reported depend on the nature of the measure that the researcher uses. The three most common types of reliability are *test-retest reliability*, *internal reliability*, and *interobserver reliability*. For example, if we argue that the SAT reflects a stable measure of scholastic aptitude in math and verbal domains, then we would expect that we would have similar scores if a person took the test today and then again next week. This type of reliability is known as *test-retest reliability*. When measures comprise multiple items that form an overall index and score of that construct, another index of reliability assesses how consistently all of the items of a measure assess the same construct. This type of reliability is known as *internal reliability* or *internal consistency*. For example, the SAT verbal score is made up of multiple items, and thus we would hope to see that the test items have high internal reliability. A test with low internal reliability suggests that participants are not responding consistently to the items and that the set of questions fails to assess the same construct. Or, if a measure is collected by having observers watch and record a behavior, another helpful type of reliability is *interobserver reliability*. Rather than having one person observe and score a behavior, researchers benefit by having multiple people observing and then assessing inter-rater reliability by comparing how similar the ratings are from different observers.

So far we have covered one major type of construct found in research, variables that are *measured*. However, researchers also are interested in a second type of construct: variables that are *manipulated*. For example, we asked you to consider the construct validity of SAT tests, and whether the SAT reflects a valid measure that should be used in college admissions. This example involves only a *measured* variable. But now consider a study in which a researcher is interested in evaluating a manipulation that could improve students' SAT scores. For instance, a researcher may be interested in formally testing whether students who receive SAT coaching before taking the SAT perform better than students who do not receive coaching. One form of SAT coaching would be buying a study guide for the SAT before taking the test. Indeed, another major facet of the Princeton Review (in addition to writing college review guides) is to provide prep guides for college entrance tests. Now, we have an example of a study with a *manipulated* variable (receiving or not receiving SAT coaching) and a *measured* variable (SAT score). Just as we are interested in evaluating the construct validity of measured variables, we are equally interested in evaluating the construct validity of manipulated variables.

One effective tool to determine if researchers have successfully manipulated a particular variable is to conduct a *manipulation check* to assess if the treatment functioned as the researchers intended. For example, if coaching entailed studying a SAT prep guide, then one possible manipulation check question could be asking whether study participants used prep guides before the exam. If we found that students in the treatment group used their prep guides while students in the no-treatment group refrained from using any prep guides, we would have construct validity evidence that our manipulation was accurately followed by participants. Alternatively, if we found that students in both groups used prep guides or received other types of coaching, then we would have evidence that our manipulation was contaminated and failed to represent the manipulated construct we intended.

In sum, construct validity entails separately evaluating whether researchers have accurately measured and/or manipulated each of their variables under investigation. If we think one or more variables are suspect in accurately representing what the researchers want represented, then we would critique the study on its overall construct validity. We have covered a lot of material, and once again it could be easy to get lost in terms and jargon. However, to help reinforce your ability to literally see the "bigger picture," look again at Figure 7.1 to see how construct

validity provides an overarching umbrella to organize and synthesize a wide range of associated concepts.

EXTERNAL VALIDITY

External validity involves making a systematic evaluation of the accuracy and strength of the ability to generalize the results beyond the current study. When assessing external validity, we typically ask if the study's results can be generalized and applied to other people, settings, or times.

Once again, many related validity terms and subtypes of external validity have been proposed that we would like to organize under the overarching umbrella of external validity (e.g., Bracht & Glass, 1968). Specifically, *population validity* entails the extent to which findings can be generalized beyond the participants of a study to the population at large. *Ecological validity* entails the extent to which findings can be generalized across different settings, particularly more naturalistic, real-world settings. *Temporal validity* entails the extent to which the findings of a particular study can be generalized across time. For example, would the results of a study conducted in 1950 still apply and generalize to today if the study was replicated? Rather than talking about external validity in broad terms, these additional concepts have appeared to help researchers highlight a specific facet of external validity focused on generalizing across people, settings, *or* time.

In addition, new external validity terms have been coined to capture unique generalizability issues in particular fields. For example, in industrial/organizational psychology, the terms *intraorganizational validity* and *interorganizational validity* (Goldstein, 1991) emphasize whether the practices found to work in one organization/company can generalize to other parts of the same organization/company (intraorganizational validity), or if findings can generalize to other organizations/companies (interorganizational validity). This should help you appreciate how validity tools evolve and are added to the research literature to help hone our ability to stress what is and isn't strong about a particular research study.

Different psychological fields have also called into question the external validity of their findings. For example, Sears (1986) wrote an influential article, titled "College Sophomores in the Laboratory: Influence of a Narrow Data Base on Social Psychology's View of Human Nature," that questioned the external validity and generalizability of social psychological research. In particular, Sears conducted a review of research published in the top social psychology journals in 1980 and 1985 and coded each study for the type of participants and the type of settings used. He documented an alarming trend: Researchers were overwhelmingly conducting studies on one particular type of participants (with over 80 percent involving college students) in one particular type of setting (with over 70 percent in an artificial, laboratory environment). This trend was in stark contrast to research conducted a quarter of a century earlier in social psychology that tested more diverse types of participants in more naturalistic, real-world settings. Sears provided compelling evidence to question the external validity of many of social psychology's findings and suggest how the unique characteristics of oversampling college students in laboratory settings may bias and limit our understanding of psychological phenomena. As a result, he challenged the field to return to conducting research that would allow stronger claims of external validity to be made.

If we return to our opening case study of UW being the No. 2 party school, how would you critique the study on its external validity? Remember, our students often report being impressed that the Princeton Review surveyed 56,000 students. However, an important conclusion of taking coursework in statistics and research methods is learning that bigger samples aren't necessarily better. One essential key to conducting research and being able to generalize back to a larger population is to ensure that we *randomly sample* participants from the population that we are interested in studying. Random sampling ensures that each participant in the targeted population has an equal chance of being included in the study. If we randomly sample, we can be more confident that we have captured a representative sample of the population and that the results based on our sample can be generalized to the population as a whole. In contrast, many researchers engage in *convenience sampling*, where the sample is simply made up of participants who are easily accessible and willing to become involved with the study. Unfortunately, convenience sampling can lead to a host of systematic biases of who is studied. Because a convenience sample is unlikely to reflect accurately all the characteristics of a given population, we are in a weak position to generalize back to the larger population.

As it turns out, the Princeton Review only used convenience sampling. They simply went to different college campuses and set up survey booths in high-traffic areas on campus (e.g., outside a student union) and asked students who were willing to fill out surveys about their college. Would students who took the time to stop to fill out the surveys accurately represent what all students on that campus thought about their college? Because the answer is clearly "no," the Princeton Review study lacks population validity. Then, regarding temporal validity, recall the student quoted in the party school article who noted that the timing of a visit (a football weekend versus the dead of a Wisconsin winter) would likely produce very different impressions of the campus. And finally, across the years, UW's No. 2 ranking has fluctuated. The Princeton Review reported UW was the No. 2 party school in 1997 and 1998, fell to No. 20 in 1999, but was back to No. 2 in 2003. Thus, to help ensure external validity, researchers need to consider carefully how to sample their participants and to consider testing a wide array of situations across time if they want to argue that their study is high in external validity. Figure 7.2 shows a summary.

Figure 7.2 Specific terms and concepts associated under the umbrella of external validity.

INTERNAL VALIDITY

Internal validity is an evaluation of a special type of "internal relationship" that can occur between variables in a study. Specifically, internal validity evaluates whether a cause-and-effect relationship between variables can be determined. Cause-and-effect conclusions hold a special place in psychology because we are ultimately interested in trying to explain the underlying cause of behavior. If a researcher does X, will it cause Y to occur? For example, if a college campus is afraid of a growing alcohol problem, can a treatment be administered on campus with the goal of decreasing the amount of drinking that occurs?

The 19th-century philosopher of science, John Stuart Mill, argued that three criteria are needed if we are to deduce that two variables are causally linked. First, if one event causes another, then the cause must precede the effect in time. This idea is called *temporal precedence*. Second, if one event causes the other, then the two events must covary together—if one changes, the other must change too. This idea is called *covariation of events*. For example, if X occurs, then Y occurs. Similarly, if X doesn't occur, then Y doesn't occur either. Third, if one event causes the other, you want to be certain that the hypothesized cause is the only cause operating in the situation to create the effect. This idea is called *elimination of plausible alternative causes*. If another variable could have created the same effect, then you cannot make a strong causal inference.

Researchers attempt to meet these criteria through experimental research designs. An experiment allows researchers to manipulate when X occurs and when Y is measured (establishing temporal precedence). An experiment has at least two groups being compared to each other, typically where one group gets the experimental treatment and another serves as a nontreatment comparison group (i.e., control group). Comparing an experimental group and a nontreatment group allows a researcher to see what happens to Y when X is manipulated (establishing covariation of events). Finally, in an experiment, researchers strive to control extraneous variables that also may cause changes in Y (ruling out plausible alternative causes). To gain this level of control, researchers often conduct their research in controlled laboratory settings rather than in more naturalistic field settings, which are far more difficult to control in order to rule out other extraneous variables. In fact, this methodological decision quickly highlights a trade-off that frequently occurs between validities: The steps that are taken to improve internal validity often limit and reduce a study's external validity.

If we return to our opening case study, how would you critique the study on its internal validity? First, internal validity is about establishing a cause-and-effect relationship between two or more variables, therefore a research study must have multiple variables under investigation and have a proposed research question that X is the cause of Y. Second, the only type of study that can establish clear cause-and-effect relationships is an experiment. If nonexperimental methods are adopted (e.g., descriptive or correlational methods), then the study is by default weak in internal validity. Finally, if an experiment is conducted, all three of Mill's criteria for establishing cause-and-effect claims need to be met.

A review of the case study quickly reveals that we are unable to meet these required conditions, and thus it is low in internal validity. For example, the case study involves only one measured variable, partying, therefore there is no proposed cause-and-effect question to be addressed. In addition, the methodology of our case study is nonexperimental. The Princeton Review simply surveyed students about their partying behavior. However, the high rates of drinking on college campuses have led many universities to engage in more experimental designs to evaluate and test the effectiveness of different types of treatments to reduce alcohol consumption and to teach students to become socially responsible drinkers (e.g., Wechsler et al., 2003). In this case, we would have multiple variables under investigation in the form of a cause-and-effect research question to test if the campus' alcohol intervention (i.e., manipulated X variable) is the cause of a reduction in alcohol consumption (i.e., measured Y variable). Figure 7.3 gives a summary picture of our overarching internal validity umbrella and related concepts to determine internal validity.

STATISTICAL CONCLUSION VALIDITY

Our final class of validity is *statistical conclusion validity*, which evaluates the accuracy and strength of the data

Figure 7.3 Specific terms and concepts associated under the umbrella of internal validity.

analyses and statistical conclusions that are drawn from a study. If you were ever wondering why psychology majors take courses on statistics, it is so we can learn how to analyze variables appropriately and correctly and draw conclusions from our research studies. In particular, we have a number of important tools to judge if a study is high in statistical conclusion validity (see Wilkinson & Task Force on Statistical Inference, 1999).

First, we can judge whether the study used the appropriate statistics to summarize and analyze the data. A researcher has many choices on which statistics to use and report. However, for a particular statistic to be appropriate it needs to meet the *statistical assumptions* that are necessary to use it.

Second, after the appropriate tests are conducted, we need to determine whether the researcher has made a *correct statistical decision* or committed a *statistical error.* Statistics allow us to make informed decisions about data. However, statistics are based on probabilities, and thus there is the possibility of arriving at a false conclusion. In particular, training in statistics will teach us how to avoid *Type 1 errors* (detecting a statistically significant effect in our data when one does not exist) and *Type 2 errors* (failing to detect a statistically significant effect when there truly is one).

Third, related to committing statistical decision errors (specifically, Type 2 errors) is another important aspect of statistical conclusion validity that involves having adequate power to test research ideas. *Statistical power* represents the degree to which a study is able to evaluate a statistical effect and to find it reliably when it truly exists. Unfortunately, many research studies can be grossly underpowered, and thus are prone to Type 2 errors.

Finally, an important component of statistical conclusion validity is moving beyond simple statistical significance tests and reporting the overall *effect size* of a study. Effect size provides a direct measure to determine if the effect of a study is small, moderate, or large in magnitude, and reflects the practical or meaningful significance of a study.

Returning one last time to our opening case study, you may find that critiquing the study on its statistical conclusion validity may be the most difficult validity of the four to judge. First, having some background with statistics is necessary before you can begin to critique the statistics of a study. Second, even with this background, popular media reports of research studies rarely provide detailed accounts of the statistical analyses used by the researchers who conducted the study. Therefore, offering a judgment about the strength of the statistical conclusion validity is extremely difficult. Instead, it is easier to offer a critique of what we would need to see before being able to make any definitive conclusions. For example, we would need to know more about the data collected by the Princeton Review to determine what assumptions were met and what statistical analyses were appropriate and inappropriate to use. Figure 7.4 provides a summary picture of our overarching umbrella for statistical conclusion validity and related validity terms and concepts.

SUMMARY

Although there are numerous terms and concepts used to describe validity, we can simplify the complexity of validity by organizing terms and concepts into four main types: *construct validity*, *external validity*, *internal validity*, and *statistical conclusion validity*. Mastering these Big 4 types of validity provides a comprehensive framework to judge the strengths and weaknesses of scientific research.

Determining the overall validity of research is simply not a *black-and-white decision.* Instead, we have four types of validity to reinforce the notion that we have many different ways to determine the strength of a research study. In fact, rarely will you ever encounter a study that ranks strong in all four validities. Remember, the methodological decisions made to increase one type of validity often harm another type of validity (e.g., steps taken to increase internal validity often harm external validity), and similar tradeoffs can occur among all four of our major classes of validity (Mitchell & Jolley, 2006).

In addition, many studies are conducted that do not address all four validities. Recall that only an experimental research methodology is able to establish internal validity and cause-and-effect conclusions. Thus, when nonexperimental research methodologies are adopted, a study by default will be weak in internal validity. Similarly, many researchers have aptly noted that external validity can be a premature goal of research (Mook, 1983). Researchers may

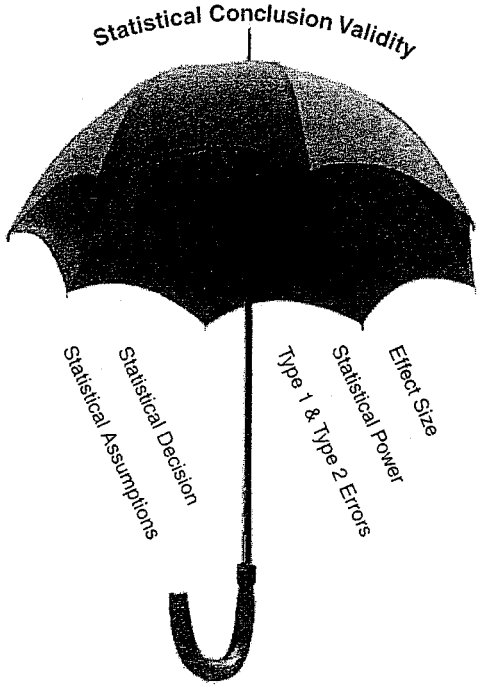

Figure 7.4 Specific validity terms and concepts associated under the umbrella of statistical conclusion validity.

first want to study a phenomenon in isolated populations or isolated situations before trying to generalize their results to the population from which they drew their sample. Therefore, to determine the overall merit of a study, we initially need to recognize what the goals of the study being conducted are and what validities need to be established to meet those goals. The next step would be to weigh the strengths that make the study valid against the weaknesses that threaten the study's validity.

Because individual studies are likely to be weak on one or more types of validity, researchers rarely conduct and report single studies. Instead, they often conduct multiple studies to answer the same research question using different research methodologies. A validity weakness threatening the conclusions of one study can be tested in a new study using a different methodology to address (or remove) that potential threat to validity. A common practice is for researchers to engage in systematic research in which a series of studies are conducted and published over time. Alternatively, researchers will report multiple studies in a single professional article to demonstrate how different threats to validity were tested and ruled out across different studies of the article.

A common area of confusion that can occur when learning about validity is determining which type of validity is being threatened. We have found that students particularly have difficulty when trying to distinguish construct validity problems from internal validity problems. Our first tip is to highlight the fundamental difference between these two ways to critique a research study. A critique of construct validity involves thinking about each variable of a study separately, and then making a judgment on whether each variable was measured or manipulated validly. If you have concerns that a particular measure or manipulation was biased in representing what it was supposed to represent, then your concern involves a critique about construct validity. In contrast, internal validity involves thinking about the relationships among variables of the study, and then making a judgment that one or more of them can be clearly linked to causing other variables in the study. Our second tip is to highlight that each of our validity concepts provides a guiding framework on how to evaluate a study. However, once again, diagnosing validity isn't always a black-and-white decision. Instead, a particular issue about a study may influence multiple validities, blurring some of the clear boundaries that fit nicely under one, and only one, validity label.

Don't be surprised when learning about validity if only *some* of our Big 4 validities are discussed. Our preference is to present validity and the Big 4 types of validity in an integrated fashion to give students a coherent picture to compare and contrast key validity concepts. However, when reviewing undergraduate psychology research methods textbooks, only half of the textbooks presented validity concepts together in an integrated fashion. Just as common was seeing different validity concepts presented in separate chapters. Moreover, although 93 percent of the textbooks discussed construct, external, and internal validity, only 40 percent discussed all four validity concepts. Also, don't be surprised if certain readings about validity focus exclusively on a particular facet of validity such as construct validity (e.g., Benson, 1998; Messick, 1995; Shepard, 1993).

Research methodology and terminology evolve over time, so you may encounter new validity terms that have been developed or validity terms specific only to certain disciplines within psychology. Try to place the new term under one of our Big 4 overarching umbrella terms to understand its purpose and how we can use it to judge a study.

We encourage you to continue to be a student of validity. We have a number of additional readings in our references that we recommend. In particular, we would like to highlight Shadish et al. (2002) as the most comprehensive resource on the Big 4 validities. We also strongly recommend taking all the coursework that you can in statistics and methods, and recommend firsthand experiences working on research projects to appreciate how each of your decisions in the research process can impact the overall validity of your conclusions. We also encourage you to practice using our Big 4 validity framework in your everyday life. You could begin by creating an exercise like the one in this chapter by picking up your own school newspaper (or, for that matter, any newspaper) and reading it critically. We also encourage you to continue using validity as a framework to judge what you can and cannot conclude from the information around you.

REFERENCES AND FURTHER READINGS

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues & Practice, 17,* 10–22.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5,* 437–474.

Bryant, F. B. (2000). Assessing the validity of measurement. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 507–619). Washington, DC: American Psychological Association.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand-McNally.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). Needham Heights, MA: Allyn & Bacon.

Goldstein, I. L. (1991). Training in work organizations. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 507–619). Palo Alto, CA: Consulting Psychologists Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387.

Rosnow, R. L., & Rosenthal, R. (1996). *Beginning behavioral research: A conceptual primer* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Sears, D. O. (1986). College sophomores in the laboratory: Influence of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Wechsler, H., Nelson, T. F., Lee, J. E., Seibring, M., Lewis C., & Keeling, R. P. (2003). Perception and reality: A national evaluation of social norms marketing interventions to reduce college students' heavy alcohol use. *Journal of Studies on Alcohol, 64,* 484–494.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 45,* 1304–1312.