aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2020, Vol. 15(2) 483–496

DOI: 10.1177/1745691619876279
www.psychologicalscience.org/PPS
SAGE

# Weighing People Rather Than Food: A Framework for Examining External Validity

**Caitlin M. Loyka, John Ruscio, Andrew B. Edelblum, Lindsey Hatch, Brittany Wetreich, and Amanda Zabel**
Department of Psychology, The College of New Jersey

**Abstract**
Research training in psychological science emphasizes common threats to internal validity, with no comparably systematic or rigorous treatment of external validity. Trade-offs between internal and external validity are well known in some areas (e.g., efficacy vs. effectiveness studies in clinical psychology), less so in others (e.g., forensic research on eyewitness identification, false memories, or confessions). We present a framework for examining external validity grounded in four domains—populations, settings, outcomes, and timeframes—that can be used to enhance the generalizability of findings. We discuss this framework and then illustrate its use by reviewing mindless eating interventions intended to help people lose weight. Research in this published literature seldom samples from appropriate populations (e.g., overweight or obese individuals) or measures appropriate outcomes (e.g., weight change) in appropriate settings (e.g., the home) over appropriate timeframes (e.g., sustained interventions with follow-up) to determine whether practical advice is empirically supported. In their applied work, we encourage psychological scientists to design studies, analyze data, and report findings with greater attention to external validity to demonstrate, rather than assume, the generalizability of findings to the intended populations, settings, outcomes, and timeframes. Editors and reviewers can hold investigators accountable for doing so.

**Keywords**
behavioral economics, health, food

Training in the research methods of psychological science emphasizes the importance of attaining strong internal validity by minimizing confounds that make it difficult to draw causal conclusions. Textbooks routinely review a fairly standard checklist of threats to internal validity (Campbell & Stanley, 1966) that includes history, maturation, testing, instrumentation, selection, statistical regression, and mortality. Research designs are then evaluated with respect to how effectively they eliminate or minimize these threats to internal validity. Experimental designs with features such as control groups and random assignment to conditions fare especially well because they do a superb job of controlling confounding influences. There is no question that internal validity is crucial in basic science, in which investigators seek to build a knowledge base consisting of causal theories used to make verifiable predictions. In the realm of applied science, in which investigators seek to develop interventions that put the basic science into practice, it becomes at least as important to establish strong external validity by providing compelling evidence that the findings will generalize in the intended ways (Steckler & McLeroy, 2009).

The authors of basic science articles themselves often speculate about applied implications of their findings, and many psychological scientists wish that behavioral research was used more often to inform public policy (Amir et al., 2005; Teachman, Norton, & Spellman, 2015), which makes it important to speak to the concerns of potential users. As Glasgow et al. (2006) put it, "The questions and concerns of clinicians, administrators, and

**Corresponding Author:**
John Ruscio, Department of Psychology, The College of New Jersey, Ewing, NJ 08628
E-mail: ruscio@tcnj.edu

policymakers are related more to external validity, generalization, and applicability of findings" (p. 106). Likewise, a number of authors have lamented how rarely basic science is translated into useful applications and discussed barriers in this process (see, e.g., a series of articles on closing the gap between research and practice; Kerner, Rimer, & Emmons, 2005). One recommendation is to pay greater attention to external validity and design studies that aim to translate science into practice.

The main goal of the present article is to raise awareness of important concerns regarding external validity that should be given serious consideration, especially in applied psychological science. We suggest adopting a framework for evaluating external validity grounded in four domains: populations, settings, outcomes, and time frames. These domains are certainly considered by researchers at times, but such consideration could be done more regularly and systematically with the adoption of a common framework. Specific criteria within each domain can be developed for a particular research topic. Having a framework for thinking about external validity to accompany the standard checklist of threats to internal validity would underscore the importance of both. The presentation of two lists of concerns would lead nicely into a discussion of their relative importance at different stages of research (i.e., basic vs. applied) and how to deal with some of the difficult trade-offs involved when designing studies.

## Framework for External Validity

The framework presented here was inspired primarily by the thoughtful discussions of generalizability in Green and Glasgow (2006) and Jenkins (1979). Green and Glasgow organized the relevant issues using a quality rating checklist with four broad domains for evaluating the external validity of public health research: reach and representativeness, program or policy implementation and adaptation, outcomes for decision making, and maintenance and institutionalization. Each of these domains is fleshed out with specific items, such as the target audience, staff expertise, moderators, costs, long-term effects, and attrition. We adopted this general approach, but we modified the domains to be more applicable to the generalizability of research in psychological science: populations, settings, outcomes, and time frames.

These domains overlap substantially with those of Jenkins's (1979) tetrahedral model of memory research, which includes subjects, instructions, materials, and outcomes. Both frameworks include outcomes, and there is a close correspondence between the domains of populations and subjects. As detailed below, the domain of settings is more broadly inclusive than that of instructions and materials, and the domain of time frames is added to cover additional features important in applied research.

To describe the four domains in our framework, we pose questions that should be given serious attention when designing or evaluating applied research, followed by illustrative questions that deal with specific concerns pertinent to the subject of particular investigations. Throughout this section, we use the generic term *treatment* to represent any type of program, intervention, or other application of psychological science designed to attain a desired outcome. These treatments could be self-administered, delivered by a professional, offered through an institution, or provided in other ways.

### *Populations*

What is the population for which the treatment is intended, and does the research obtain a sample representative of that population? This domain is similar to the domains of reach and representativeness in Green and Glasgow (2006) or subjects in Jenkins (1979). Importantly, we have in mind much more than demographic information about the intended population. For example, is the treatment designed to improve psychological functioning for people experiencing distress or impairment or prevent signs or symptoms among those that are currently mentally healthy? Is the treatment aimed at differentiating between guilt and innocence among criminal suspects? Is the treatment meant for use with children, adults, or both? These questions are just some that should be considered when conceptualizing the intended population. Once the intended population has been established, the next step is to ask whether care has been taken to ensure representative sampling. Was a convenience sample used, or were research subjects randomly selected from the intended population? Does volunteer bias or attrition threaten the representativeness of the sample?

### *Settings*

What is the setting in which the treatment is intended to take place, and is the research conducted in a representative setting? This domain overlaps somewhat with the domain of program or policy implementation and adaptation in Green and Glasgow (2006) and includes the instructions and materials domains in Jenkins (1979) as well as some additional issues. For example, is the treatment meant to be self-administered at home or delivered by a professional in an office? If the treatment is designed to help people break bad habits or develop better habits (e.g., eating a more healthful diet, exercising more, quitting smoking), what

are the contexts in which people grapple with these changes? Does the setting include other people, and if so, are they friends, acquaintances, professional colleagues, or family? Once the intended setting has been established, the next step is to ask whether care has been taken to ensure a representative research setting. Was the study performed in a laboratory or in a more naturalistic environment appropriate to the intended setting? How closely do behavior, stimuli, or decisions in the study mimic those in real life as opposed to being hypothetical? If there is something significant at stake, does the study capture it or simulate it with a low-stakes alternative? Do aspects of the social environment correspond to what is intended, or was the study conducted in relative isolation or with strangers?

### *Outcomes*

What are the most important outcomes the treatment is intended to achieve, and does the research measure these? This domain is very similar to the corresponding outcomes domains in Green and Glasgow (2006) and Jenkins (1979). For example, is the treatment designed to change attitudes, cognition, or behavior? Is the hope that the treatment will cure a condition or reduce its severity to some degree? Is the treatment meant to affect a single important decision or promote long-term behavior change? Will the outcomes of interest be experienced by the treated individuals, significant others in their lives, their employers, or members of an entire community? Once the most important outcomes have been established, the next step is to ask whether care has been taken to measure them effectively. Do measures correspond to important real-world goals, or are more convenient proxy variables assessed? Were outcomes assessed using self-report, informants, or observation, or were they recorded more objectively through physiological or other means? Do outcomes include measures pertinent to costs or potential adverse consequences? Do measures afford tests for moderator effects or sensitivity analyses to examine the robustness of treatment effects? Are outcomes reported in ways that enable the target audience to make informed decisions about treatment effectiveness?

### *Time frames*

There are two distinct facets to the final domain: What is the intended duration of the treatment, and when should outcomes be assessed? These issues are not mentioned by Jenkins (1979), but they do overlap with those in Green and Glasgow's (2006) domain of maintenance and institutionalization. For example, is the treatment designed to be a single experience, or will it entail repeated exposures to stimuli or conditions? Is the treatment meant to be sustainable beyond the period during which it is introduced? Are effects expected immediately, or are the most important outcomes intended to emerge over time? Once observed, should the effects dissipate or persist? Once the time frames have been established, the next step is to ask whether care has been taken to design research such that the treatment and assessment plans correspond to the goals. Does the duration of treatment in the research approximate what is intended in practice? If an abbreviated treatment is administered, what steps have been taken to determine whether a novel stimulus produces results that would dissipate through habituation or whether further exposure might reinforce the treatment and strengthen an initially weak effect? Are outcomes measured at appropriate time points? If there is no follow-up or it spans a briefer time than effects are meant to last, what assurance can be provided that effects would not diminish over a more appropriate time horizon?

## An Illustrative Review: The External Validity of Mindless-Eating Research

The previous section provided an overview of our proposed framework for external validity. Not only can this be used in research training and study development, but also as Glasgow and colleagues have both recommended and demonstrated (Glasgow et al., 2006; Green & Glasgow, 2006; Klesges, Dzewaltowski, & Glasgow, 2008), external validity can itself be reviewed systematically. Our proposed framework provides four domains to consider, and specific issues pertinent to a research area can be fleshed out to develop a coding scheme to assess the extent to which generalizability to the intended populations, settings, outcomes, and time frames has been empirically tested. An even more stringent approach would ask whether a body of literature has successfully passed such empirical tests. To illustrate how to do so using our proposed framework, we performed a review of the literature on eating interventions intended to help people achieve or maintain a healthy weight.

Wansink (2006, 2014) referred to the "mindless-eating" habits that might cause problems with weight management and discussed a wide range of innovative ideas, cleverly designed studies, and compelling findings from his lab and related behavioral economic research. For example, compared with students eating tomato soup from an ordinary bowl, students whose bowl was surreptitiously rigged to automatically refill itself ate more without realizing it (Wansink, Painter, & North, 2005). This finding suggests that eating cessation is influenced by external cues (e.g., how much food is

left) as well as internal cues (e.g., satiety). By understanding influences such as these, researchers have proposed many interventions to decrease mindless food intake. These interventions include reducing portion sizes, going without a tray in cafeterias, posting calorie or other nutritional information in restaurants, and shrinking container sizes.

Many of these mindless-eating interventions (Wansink et al., 2005) are based on the idea that people are generally unaware of the decisions they are making when eating and therefore try to harness the automatic and relatively effortless mode of thinking that Kahneman (2011) called System 1 (e.g., deconveniencing tempting foods, labeling food more descriptively, making serving sizes clearer). Other interventions target what Kahneman called System 2, a more reflective, deliberate, and effortful mode of thinking. For example, posting calories in restaurants or removing trays in cafeterias is meant to make consumers more consciously aware of what they are eating so they can actively make an informed choice.

Wansink's (2006, 2014) engaging books are representative of the behavioral economic approach to offering advice to individuals, businesses, or policymakers. The purported goal is to improve health, primarily by helping people to lose weight, but there are several causes for concern regarding the external validity of the studies cited in support of the practical advice. For example, one might question the populations (e.g., do results for unselected convenience samples of college students generalize to overweight or obese adults seeking to lose weight?), settings (e.g., do findings from lab settings with strangers generalize to the home or other familiar settings with family or friends?), outcomes (e.g., does the amount of food eaten in a single sitting have much to do with weight change over time, rather than being partially or fully offset by unstudied behavioral compensation?), and time frames (e.g., does the response to a novel stimulus persist with repeated exposures or dissipate through habituation?) involved in this research.

Failing to address concerns such as these necessitates extrapolations based on questionable assumptions to make the case that mindless-eating interventions would help individuals achieve important weight goals (e.g., achieving or maintaining a healthy weight) in the contexts of their real lives. For example, many studies in this literature record the amount of food consumed, weighing this rather than the participants. However, the food consumed in one exposure to an intervention is a very poor proxy for the real goal, weight loss. Suppose that a study finds people consume 100 fewer calories in an experimental condition relative to a control condition. A researcher might then extrapolate from these results, assuming that if a person does this each day, this would lead to a gain of 1 pound every 35 days, or about 10 pounds per year. This line of reasoning is the basis for most of the advice in Wansink's (2006) book on mindless eating, given that he suggested that people can make small changes along a "mindless margin" that will lead to substantial weight loss over time. The opening chapter is very explicit about this assumption, repeatedly insisting that small changes in behavior will have long-term consequences that can be predicted in a straightforward manner. For example, Wansink wrote that "Just 10 extra calories a day—one stick of Doublemint gum or three small Jelly Belly jelly beans—will make you a pound more portly one year from today" (p. 31).

This assumption, however, is not as plausible at it might seem. A large, interdisciplinary team of specialists published a review of myths, presumptions, and facts about obesity in a leading medical journal (Casazza et al., 2013), and the first of seven myths was the following: "Small sustained changes in energy intake or expenditure will produce large, long-term weight changes" (p. 447). One minor, technical reason for the inaccuracy of this is that an initial gain in weight increases one's daily caloric needs, so a simple linear extrapolation that fails to adjust for this will fail. More important, however, are empirical findings that such extrapolations predict greater weight change than is observed. It appears that homeostatic mechanisms, such as speeding or slowing metabolic rates to restore energy balance, or behavioral compensation (e.g., while consuming extra calories of a food or drink, consuming fewer calories of other items at the same sitting, elsewhere in their day, or over time) may mute the impact of small changes in caloric intake or energy expenditure. The dubious nature of a key assumption permeating the research on mindless eating underscores the importance of testing the generalizability of findings in this literature. Our illustrative review will examine the extent to which a number of concerns regarding external validity have been addressed in research on interventions related to mindless eating.

## Method

### *Sampling procedure*

A literature search was performed to identify articles that might qualify for inclusion in the review. To cast as wide a net as possible, subject matter spanned research on eating habits, eating interventions, weight management, and eating behaviors with applied implications. Topics included interventions entailing proximity and visibility of food; container, utensil, and plate size, shape, and color; portion size; packaging design; menu descriptions and nutrition labeling; price signals; segmenting food; tray-less cafeterias; variety of food

choice; distractions present during consumption; social influence; and atmosphere of the eating environment. These topics were divided among four researchers for the initial article search. Candidate studies were found using the PsycINFO, PubMed, and Google Scholar databases in 2016. Searches began with the following keywords drawn from familiarity with the literature, which were sometimes cross-referenced to narrow an overwhelmingly large set of results: *atmosphere; calorie consumption; container, utensil, or plate size, shape, or color; distractions; menu descriptions; nutrition labels; packaging design; portion size; proximity; price signals; segmenting food; social influence; trayless cafeteria; unit bias; variety; visibility*. Perhaps the most important and fruitful part of the process was that after using titles and abstracts to identify potentially relevant articles, further searches were performed both backward (by scanning all items in each candidate article's reference list) and forward (by retrieving lists of all publications that had cited each candidate article, using PsycINFO) to identify other articles that might meet inclusion criteria.

This process yielded a total of 188 articles to scan for studies that satisfied six inclusion criteria: The articles had to be peer reviewed, be written in English, have adult subjects, had to have potential applications of findings, have empirical comparisons, and have physical or behavioral outcomes. Peer-reviewed journals ensured minimal quality standards. Publication in English ensured that we could accurately code the research. Requiring that subjects be at least 18 years old ensured that we reviewed research with implications for adults. Research with children often served very different purposes, with different audiences in mind (e.g., schools rather than individuals or their families). Unless the authors explicitly addressed the potential application of their findings, a study was excluded on the grounds that it was basic rather than applied science. Requiring empirical comparisons (e.g., experimental vs. control conditions) eliminated purely descriptive studies. Conditions could be manipulated or measured, and within-subject comparisons were allowed so that longitudinal research could qualify. Outcomes had to include at least one physical measure (e.g., weight, body mass index, or change in either) or behavioral measure (e.g., amount of food eaten, type of food chosen). Attitudes, beliefs, or stated intentions were not counted as behavioral outcome measures. A total of 156 studies, appearing in 121 articles, met all six inclusion criteria. References for these articles appear in the Supplemental Material available online.

Although we developed our own array of topics to begin the review, articles were subsequently categorized using a two-tiered system presented in Wansink and Chandon (2014). The higher-order categories are factors relevant to food consumption monitoring: sensory cues, emotional cues, and normative cues. Sensory cues include the four subcategories of hunger and satiation cues (determined by energy inputs and outputs); palatability (based on the sensory properties of food); ambient sound, scent, lighting, and temperature; and individual differences such as cognitive restraint, the distinction between restrained and unrestrained eaters. Emotional cues include the two subcategories of affect valence (positive or negative emotions, goal-dependence of emotions) and stress. Normative cues include the three subcategories of social facilitation (social cues and matching of appropriate food intake quantities), categorization cues and health halos (influenced by perceived healthfulness of food), and portion size cues (determined by packaging and the size of dinnerware). This two-tiered classification is detailed in Appendix A, and Table 1 shows how many studies were placed into each category.

A final note on sampling is unusual but necessary in this case. When we learned that questions had been raised about the integrity of research performed in Dr. Brian Wansink's Food and Brand Lab at Cornell University and later that allegations of scientific misconduct had been made, we decided to put this project on hold until it became clear which, if any, of the studies in our review might be retracted or corrected. We waited until Cornell had completed its investigation in September 2018 and the dust appeared to have settled on the intense scrutiny of research connected to this lab. Dr. Wansink was an author on 14 of the 121 articles in our review (12%), which contained 21 of the 156 studies

**Table 1.** Number of Qualified Studies in Each Topic Category

| Topic category | Number of Studies |
|---|---|
| Sensory cues | |
| Hunger and satiation cues | 22 |
| Palatability | 18 |
| Ambient sound, scent, lighting and temperature | 10 |
| Individual differences and role of cognitive restraint | 5 |
| Emotional cues | |
| Affect valence | 0 |
| Stress | 2 |
| Normative cues | |
| Social facilitation and matching | 27 |
| Categorization cues and health halos | 35 |
| Portion size cues | 35 |
| Total | 154 |

Note: Two of the 156 studies could not be categorized using this classification.

(13%). As of this writing (July 29, 2019), none of the articles cited in the text or included in our review have been retracted or corrected. The allegations and misconduct appear to center around issues such as *p*-hacking and false-positive findings (Simmons, Nelson, & Simonsohn, 2011), which are most likely to entail concerns about the conclusions drawn in these studies and not about issues of study design that would affect external validity. In fact, removing all studies involving Dr. Wansink would have very little effect on any of the results, and no effect on the conclusions, from our review.

## *External validity criteria*

Our framework was used to create a set of criteria for coding studies. The four domains of populations, settings, outcomes, and time frames each contained specific criteria. Coders recorded whether each criterion was met and took more detailed explanatory notes. Appendix B describes the coding scheme.

In the populations domain, the first criterion was whether the authors explicitly stated the intended population. The second criterion was whether a sampling plan was used to achieve representativeness of the sample to the intended population (e.g., random sampling). Convenience sampling would only qualify if inclusion or exclusion criteria were used to attain a representative sample. The third criterion was whether there was implicit or explicit assurance that attrition did not threaten the final sample's representativeness to the intended population.

In the settings domain, two criteria were similar to criteria in the populations domain. The first criterion was whether the authors explicitly stated the intended setting. The second criterion was whether steps were taken to ensure the representativeness of the study setting to the intended setting.

In the outcomes domain, the first criterion was whether measures constituted the outcomes of genuine interest (e.g., weight of subjects) rather than proxies (e.g., amount of food consumed). The second criterion was whether the authors reported findings in ways that address important goals (e.g., reaching or maintaining a healthy weight). The third criterion was whether the authors considered any kinds of behavioral compensation or ways that effects observed in the study could have been partially or fully offset by behaviors beyond those observed during the study itself (e.g., adaptation over time, eating less after the study session). The fourth criterion was whether the authors addressed adverse consequences of the intervention (e.g., inconvenience, negative emotions, reduced quality of life). The fifth criterion was whether moderator effects were examined to assess the robustness versus specificity of effects (e.g., healthy vs. unhealthy participants, age differences, cultural differences). The sixth criterion was whether sensitivity analyses were performed to test for patterns such as dose-response effects, threshold levels, or diminishing returns. The seventh criterion was whether the authors evaluated the monetary, time, or other costs of putting proposed interventions into practice.

In the time-frames domain, the first criterion was whether the duration of the research was sufficient (e.g., repeated exposures) or extrapolation from the findings would be required to draw conclusions (e.g., a one-off observation). The second criterion was whether the authors reported data on longer-term effects following an intervention. The third criterion was whether data analysis or discussion evaluated the sustainability of an intervention (e.g., feasibility of continuing an intervention beyond the study period).

## *Reliability checks*

A team of four researchers coded studies using the criteria listed previously, and the reliability of coding was assessed by assigning studies to pairs of researchers. Approximately one sixth of the articles was assigned to each of the six possible pairings among the researchers. Coding was compared within pairs to record the levels of agreement on the external validity criteria, and these agreement levels were averaged across all pairs. Disagreements were discussed and resolved within pairs to determine the final coding for all studies. As shown in Table 2 (first column), there was generally high agreement on coding (median = 88% agreement). Seven of the 15 criteria were coded with at least 90% agreement, and only 3 of the 15 criteria were coded with less than 83% agreement.

## *Coding publication age and scholarly influence*

In addition to examining how often the external validity criteria were satisfied, we wanted to test for change over time and a relationship with influence in the scholarly literature. To do this, we coded two additional measures, which were last updated in February 2019. The first variable was publication age, recorded as 2019 – year of publication. Because this variable was skewed, we performed a log transformation. The second variable was scholarly influence, initially recorded as the citation count provided by Google Scholar. To control for skew plus the fact that studies have more opportunity for citations as they age, data were transformed as follows: influence = log10 ((citations / (2019 – year of publication)) + 1).

**Table 2.** Coding Agreement and Findings for External Validity Criteria

| External validity categories and criteria | Reliability (% agreement) | Criterion (% meeting criterion) | Correlation With Time | Correlation With Influence |
|---|---|---|---|---|
| Populations | | | | |
| Population | 65 | 47 | −.13 | .13 |
| Sampling | 70 | 44 | −.08 | −.03 |
| Attrition | 98 | 99 | −.07 | −.04 |
| Settings | | | | |
| Setting | 90 | 38 | −.06 | .08 |
| Representativeness | 87 | 38 | −.01 | .09 |
| Outcomes | | | | |
| Measures | 100 | 3 | −.16 | .00 |
| Reporting | 88 | 77 | −.07 | .25* |
| Compensation | 83 | 65 | .06 | .01 |
| Adverse consequences | 94 | 3 | −.05 | .03 |
| Moderators | 87 | 61 | −.02 | −.09 |
| Sensitivity | 99 | 3 | .06 | −.10 |
| Costs | 88 | 15 | .07 | .07 |
| Time frames | | | | |
| Duration | 92 | 43 | .17 | .00 |
| Long-term effects | 98 | 4 | −.05 | .03 |
| Sustainability | 60 | 64 | .02 | .29* |
| Percentage of criteria met | | | −.03 | .18 |

Note: Reliability was calculated for the 143 studies in the original database. Time was calculated as $\log_{10}$ (2019 – year of publication). Influence was calculated as $\log_{10}$((citations / (2019 – year of publication)) + 1). Statistical significance required $p < .0025$ because of a Bonferroni correction for multiple testing ($\alpha = .05$ divided by 20 = .0025).
*$p < .0025$.

## Results and Discussion

A total of 156 studies appearing in 121 articles met all six criteria for inclusion in this review. Results are presented using the study as the unit of analysis, but they were extremely similar when weights were applied so that the article became the unit of analysis (e.g., assigning weights of one third to the codes for each of three studies appearing in the same article). The central question motivating this review was how often the external validity criteria would be satisfied, and these findings are reported in Table 2 (second column). For interested readers, we also break these results down by topic areas in Table 3. We hesitate to draw conclusions about differences across topic areas because of the fairly small sample sizes for most of them. Particularly when the more heavily populated topic areas are considered, the trends described below appear to hold.

The intended population was stated explicitly in nearly half (47%) of the studies, and about the same proportion (44%) used a sampling method that verified the representativeness of the chosen sample to the intended population using either inclusion or exclusion criteria or random sampling. Nearly all of the studies (99%) provided assurance that attrition did not pose its own problems with respect to the representativeness of the sample. Despite the explicit or implicit goal of helping people lose weight, investigators almost never deliberately sampled from a population in need of help: overweight or obese individuals.

Just over one third of studies explicitly stated the intended setting (38%) or ensured representativeness of the studied setting to the intended setting through their design and procedure (38%). The research setting was usually a cafeteria or a lab, and in the latter case, the lab was often designed to simulate a restaurant. In addition, with few exceptions, the research settings involved situations in which participants were relative or complete strangers. Unfortunately, this literature sheds very little light on how people behave in more familiar settings. Even a simulated restaurant is very different from eating at home, at a favorite restaurant, or in any other familiar location shared with family or friends.

Virtually none of the studies (3%) measured anything other than a proxy variable. It is striking how meticulously investigators weigh the food that people have consumed but almost never the people themselves. As

**Table 3.** Percentage of Studies Meeting External Validity Criteria by Topic Category

| | Sensory cues | | | | Emotional cues | Normative cues | | |
|---|---|---|---|---|---|---|---|---|
| External validity categories and criteria | Hunger and satiation cues ($n$ = 22) | Palatability ($n$ = 18) | Ambient sound, scent, lighting, and temperature ($n$ = 10) | Individual differences ($n$ = 5) | Stress ($n$ = 2) | Social facilitation ($n$ = 27) | Categorization cues and health halos ($n$ = 35) | Portion size ($n$ = 35) |
| Populations | | | | | | | | |
| Population | 45 | 33 | 40 | 60 | 0 | 37 | 66 | 49 |
| Sampling | 50 | 39 | 50 | 60 | 0 | 48 | 49 | 37 |
| Attrition | 100 | 100 | 100 | 100 | 100 | 100 | 97 | 100 |
| Settings | | | | | | | | |
| Setting | 18 | 44 | 60 | 0 | 50 | 48 | 49 | 29 |
| Representativeness | 23 | 44 | 70 | 20 | 50 | 52 | 43 | 23 |
| Outcomes | | | | | | | | |
| Measures | 9 | 0 | 0 | 20 | 0 | 0 | 0 | 3 |
| Reporting | 73 | 78 | 70 | 80 | 0 | 63 | 77 | 94 |
| Compensation | 64 | 61 | 80 | 60 | 100 | 59 | 71 | 66 |
| Adverse consequences | 14 | 0 | 0 | 20 | 0 | 0 | 3 | 0 |
| Moderators | 45 | 72 | 30 | 20 | 100 | 70 | 71 | 63 |
| Sensitivity | 5 | 0 | 0 | 0 | 0 | 7 | 3 | 0 |
| Costs | 9 | 6 | 40 | 0 | 0 | 30 | 17 | 9 |
| Time frames | | | | | | | | |
| Duration | 68 | 44 | 40 | 60 | 50 | 19 | 54 | 34 |
| Long-term effects | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 3 |
| Sustainability | 68 | 39 | 40 | 40 | 0 | 44 | 80 | 91 |
| Criteria met (%) | 39 | 37 | 41 | 36 | 30 | 39 | 46 | 40 |

Note: Values are $n$s unless otherwise noted. No studies fell into the affect valence category.

a consequence, almost nothing is known about whether the mechanisms under study, the advice being offered, or the interventions being recommended would have practically significant effects on people's weight, let alone their health or quality of life. About three in four studies (77%) reported their findings in ways that address goals such as changing eating or purchasing behavior or reducing weight. Roughly two thirds of the studies (65%) mentioned compensation mechanisms that could partially or fully offset observed findings, such as eating less on another occasion to offset increased consumption in the lab session, although this was usually not tested empirically.

Very few studies (3%) discussed any potential adverse consequences, and there is always the risk that even a well-intentioned intervention might have unintended effects—or even backfire. More than half of the studies (61%) tested moderators but somewhat superficially. For example, the moderators typically examined were demographic variables such as gender or age but not health status or weight. It would be extremely informative to know whether an influence on eating behavior is associated with health status or weight because that might suggest it played a role in leading to obesity. Likewise, it would be important to know whether a proposed intervention to reduce food intake is effective among those who stand to benefit from weight loss. Although it might be interesting to test for differences along demographic lines, the value of such analysis is lessened considerably when the overall results are already of questionable generalizability due to serious threats to external validity. For example, it seems comparatively unimportant to know whether men or women ordered more grams of French fries in an experiment when the study itself cannot speak to any effects on outcomes of greater interest, such as weight loss.

Very few of the studies (3%) reported sensitivity analyses (e.g., testing for dose-response relationships or threshold effects), and few (15%) discussed any costs of proposed interventions (e.g., money, time, effort, inconvenience, negative emotions). For instance, mandating calorie information on menus is expensive and can induce feelings of guilt or shame when people order a food they enjoy, so one would need to judge whether any benefits outweigh these costs.

Less than one half of the studies (43%) examined a phenomenon for a duration of time that was sufficient to avoid extrapolation. Most observed behavior in a single encounter rather than with repeated exposures, and very few studies (4%) tested long-term effects. Thus, very little is known about whether people will follow through with an intervention or how they might adjust to it. It is unclear whether a response to a novel stimulus (e.g., a change in container or plate size) would dissipate through habituation if put into practice on a regular basis. Even if a response does persist, the assumption that small, sustained changes in caloric consumption will produce large, long-term weight changes remains untested. The fact that this ides has been classified as a myth of obesity (Casazza et al., 2013) suggests that advice founded on this assumption should be treated skeptically until supportive evidence is provided. About two thirds of the studies (64%) considered the sustainability of their findings after the conclusion of their research, although this was seldom tested.

In addition to examining how often the external validity criteria were satisfied, we performed two series of correlational analyses. The first series of analyses tested for change over time to determine whether external validity is being given more or less attention in published research. The second series of analyses tested for influence in the scholarly literature to determine whether external validity leads studies to be cited more or less often. Specifically, we correlated the satisfaction of each external validity criterion with the publication age and scholarly influence variables. All correlations are reported in Table 2 (final two columns). A Bonferroni correction was used to control the Type I error rate for each series of tests, meaning that the threshold for statistical significance was reduced from .05 to .05 / 20 = .0025.

The overall percentage of external validity satisfied was not correlated with time, $r(154) = -.03$, $p = .675$, or influence, $r = .18$, $p = .028$. This finding suggests that attention to external validity has changed little over time and is unrelated to scholarly influence. Among the 15 specific items on the external validity checklist, none were statistically significantly correlated with time, but two were correlated with influence (reporting, $r = .25$, $p = .002$; sustainability, $r = .29$, $p < .001$). Given the relatively small size of these correlations and the fact that there was no clear trend across external validity criteria, we are not inclined to draw any strong conclusions from these findings.

## Conclusions

This illustrative review raises serious questions about the generalizability of findings from the literature on mindless-eating interventions. At the same time, there are a few caveats to bear in mind, most of which extend well beyond this particular research area. First, publication biases may contribute to a disappointing level of attention to external validity. With a realistic understanding that the peer-review process demands strong internal validity, investigators who make design choices that strengthen internal validity at the expense of external validity may be more successful in publishing their

work than researchers who make the design choices favoring external validity. In addition, more externally valid studies may be less likely to yield statistically significant results than less externally valid ones. For example, it may be easier to attain significant findings using a proxy measure with a single exposure to a novel stimulus (e.g., amount of soup consumed from a refilling bowl in one sitting) than using better outcome measures over repeated exposures to more natural stimuli with long-term follow-up (e.g., sustained weight loss after a year of visiting a tray-less cafeteria).

For our illustrative review, studies were drawn from published research as a means of quality control. This process may have led to an underrepresentation of studies that are relatively strong in external validity. It would be interesting to know how many such studies were considered but abandoned at the design stage out of a concern that they might not be publishable or were performed but abandoned because they did not yield statistically significant results and languish in the proverbial file drawer. What they say about the effectiveness of mindless-eating interventions is unknown.

Another factor to consider in drawing conclusions from the frequency with which studies address concerns regarding external validity is that it would be unreasonable to expect every study to address all of them. Space limitations make it challenging to report all of this information in a journal article. Perhaps even more significant is the pressure to strengthen internal validity, often at the expense of external validity, and document how this was done to meet the demands of the peer-review process and get published.

Even with these caveats in mind, when an entire research literature fails to address important concerns, this should give pause to anyone citing its findings as support for advice or interventions. In the case at hand, it seems crucial that at least some mindless-eating research be done using appropriate populations (e.g., those who would like to lose weight), measuring appropriate outcomes (e.g., weight change), in appropriate settings (e.g., interventions in the home), and over appropriate time frames (e.g., sustained interventions with follow-up measures) to determine which applications are empirically supported. Comparison groups of individuals not exposed to the interventions being tested would afford some measure of assurance that causal conclusions might be justified, particularly if participants are randomly assigned to conditions. Testing proposed interventions in realistic contexts could reveal which of the intriguing mechanisms uncovered in the basic science are worth putting into practice, for whom, in what ways, and in what settings.

## *Internal and external validity*

According to Coolican and Flanagan (2005), the primary difference between internal and external validity is that internal validity involves "the need for control" and external validity involves "the need to preserve the essence of the phenomenon under investigation" (p. 24). When conducting a study, researchers must choose which type of validity they find more important, and there is often a trade-off associated with either choice. These trade-offs are well understood in fields of study centered on health promotion (Prohaska & Etkin, 2010), in which the distinction between efficacy studies (which focus on strong internal validity) and effectiveness studies (which focus on strong external validity) is the source of thoughtful discussion and debate (e.g., Clarke, 1995; Flay, 1986). In an efficacy study, subjects must meet stringent eligibility criteria (e.g., being diagnosed with a single mental disorder), therapy is delivered in a controlled manner (e.g., from a treatment manual), and data are analyzed only for subjects who complete all therapy sessions. Exerting experimental control in these ways improves the ability to draw conclusions regarding cause and effect. At the same time, this choice limits the generalizability of those conclusions to clinical practice.

Effectiveness studies are quite different. Eligibility criteria are less stringent, which allows a more representative sample of patients to be studied. Therapy is delivered in a more natural manner, better reflecting the personalization of treatment in practice. So-called intention-to-treat analyses examine data for all patients enrolled in the study, which includes those who chose to discontinue treatment for any reason. In all of these ways, effectiveness studies make it more difficult to draw causal conclusions, but they do make it easier to generalize the results to clinical practice.

Each of the four domains within our proposed framework for external validity highlights a type of trade-off with internal validity. Although there is often tension between the two, some design strategies can strengthen both. For example, if the goal of applied research is to help people lose weight, we recommend weighing people rather than the food they consume. Doing so would greatly improve external validity by measuring a more important outcome, with no cost in terms of internal validity. Likewise, studying change in weight would require a sufficient duration of study to allow nontrivial changes to occur, and doing so would not jeopardize the internal validity of a treatment study. Indeed, these changes might even strengthen internal validity. When assessing only the immediate impact of a single exposure to a novel stimulus, some or all of

an observed effect across experimental conditions might be due to reactivity or demand characteristics generated by being observed under unusual circumstances. Testing an intervention over a more appropriate duration using more appropriate outcome measures can help to determine whether there is in fact a causal relationship between the constructs under investigation rather than merely an experimental artifact. Thus, although we recognize that there will often be a tension involved in choosing between research strategies that pose trade-offs of internal and external validity, this is not necessarily the case, and there may be alternatives that strengthen both types of validity.

## *Research planning and evaluation*

Routinely thinking about the correspondence between the studied and intended populations, settings, outcomes, and time frames might help to ensure that findings will actually demonstrate generalizability or whether untested assumptions are involved. This practice should be most helpful at the research-planning stage, although more careful evaluation during the peer-review process could also be helpful to increase accountability in applied research. The scientific publication process tends to favor research with strong internal validity at the expense of external validity.

When submissions are reviewed for publication, editors rely heavily on threats to internal validity as grounds for rejection. Even when authors speculate about applications of their findings, they run very little risk if they either ignore or only superficially speak to the external validity of their findings. A thoughtful discussion of external validity will not compensate for a serious threat to internal validity. Particularly at a journal of applied psychology, or when authors of basic science choose to discuss applied implications of their findings, editors and reviewers can expect and demand a thoughtful consideration of issues related to external validity.

Simons, Shoda, and Lindsay (2017) proposed that all empirical research reports should contain a statement of constraints on generality (COG) that identifies the target population for the findings. They argued that doing this would aid attempts to replicate the findings or test their boundary conditions. We applaud this attempt to draw attention to serious concerns regarding external validity, but we believe it is both too narrow and too broad. The call for COG statements is too narrow in that it focuses exclusively on populations, and we believe it would be useful to deal with settings, outcomes, and time frames as well. It is important to consider all four domains when thinking about the generality of findings. The call for COG statements is too broad in that it is directed at all empirical research reports, not just those that deal with applied science. We suggest, instead, that investigators should be expected to discuss all four domains in our framework for external validity whenever they venture into the realm of applied science, either by submitting their work to a journal with an applied emphasis or offering speculations in their article about applied implications of their findings. An applied journal could establish its own call for COG statements, ideally requiring discussion of all four domains. Whether a journal requires such a statement, when an article deals with applied psychological science, the editor and reviewers should evaluate potential threats to external validity with the same degree of critical thinking that they devote to potential threats to internal validity.

On a related note, literature reviews, including but not limited to meta-analyses, could do a better job of dealing with external validity when the subject matter has applied implications. Threats to internal validity are dealt with through inclusion or exclusion criteria when selecting studies that qualify for review. Studies with the strongest external validity may be exposed to more threats to internal validity and thereby less likely to qualify for inclusion. To the extent that authors of review articles address external validity at all, they are more likely to note the aggregate number of subjects or anecdotally list the types of populations, settings, outcomes, and time frames among studies included in the review than to make these focal points of the investigation. A more systematic and rigorous handling of both internal and external validity would be helpful (Jüni, Altman, & Egger, 2001). It is not uncommon to rate the methodological rigor of studies included in reviews to test for moderating effects. Something similar could be done for external validity. Studies could be included even when they were designed with stronger external than internal validity. Then, our proposed framework could be used to develop a coding scheme to rate the studies' external validity. In addition to enabling further tests of moderating effects, descriptive analyses of external validity such as those presented in our illustrative review could shed considerable light on the generality of findings in a research literature.

## *Concluding thoughts*

In many areas of applied psychological science, promising findings have emerged from research that is generally strong in internal validity. In some of these areas, such as clinical psychology, practitioners are well aware of the concerns with generalizing from science to practice.

In other areas, however, it would be worthwhile to review the literature to examine the extent to which it has addressed such concerns (e.g., topics in forensic psychology such as eyewitness identification, polygraph testing, false memories, or confessions). Adopting a framework for thinking about external validity with respect to the intended populations, settings, outcomes, and time frames would facilitate such review. These domains can be fleshed out with specific criteria applicable to a particular application.

We believe this framework holds great potential as a tool for research training. Having a teaching tool to parallel the ubiquitous treatment of threats to internal validity might reinforce the importance of external validity. Moreover, an external validity framework could be used in the evaluation of applied research proposals (e.g., by thesis/dissertation committees or funding sources). If science is to be put into practice, it needs to meet the needs of its users. This entails an ability and willingness to design studies, analyze data, and report findings with a greater emphasis on external validity to demonstrate, rather than assume, the generalizability of findings to the intended populations, settings, outcomes, and time frames.

## Appendix A: Categorization of Eating Interventions

### *I. Sensory cues: how senses react to stimuli*

A. *Hunger and satiation cues*: Were variables defined in terms of energy inputs and outputs, a disconnect between hunger and amount consumed, or the memory of consumption influencing satiety cues?

B. *Palatability*: Were variables defined in terms of anticipated and experienced pleasure of eating or smelling food, the distance and appearance of food, the texture, temperature, and viscosity of food or the sound of food when shown, served, or eaten?

C. *Ambient sound, scent, lighting, and temperature*: Were variables defined in terms of external factors that cannot be blocked out, controlled, or avoided, such as background music, complementing odors, or harsh lighting?

D. *Individual differences and role of cognitive restraint*: Were variables defined in terms of individual characteristics of participants, such as their weight or body type (obese vs. non-obese), dieting status (restrained eating vs. unrestrained eating), mood, or cognitive load?

### *II. Emotional cues: feelings and attitudes*

A. *Affect valence*: Were variables defined in terms of emotions (positive vs. negative), temporal orientation and function, or goal-dependence of emotions?

B. *Stress*: Were variables defined in terms of physical stressors (including threats of shock, viewing unpleasant videos, task failures, anticipated public speaking, interpersonal rejection, remembering negative personal events, or depression)?

### *III. Normative cues: how you believe you are supposed to eat*

A. *Social facilitation and matching*: Were variables defined in terms of the impact of social cues on self-reported hunger, arousal, emotionality, duration of eating, or amount eaten?

B. *Categorization cues and health halos*: Were variables defined in terms of healthfulness claims (healthy vs. unhealthy), intrinsic quality of food items (good vs. bad), food type, brand, packaging, price, promotion and distribution, or these factors' operation (whether or not it is independent of individuals' BMI, gender, or level of restraint when eating)?

C. *Portion size cues*: Were variables defined in terms of food packaging, the amount of food left after portion has been served, or the size of dinnerware?

## Appendix B: External Validity Criteria

### *I. Populations*

A. *Population*: Did the authors state the intended population (e.g., all adults, obese individuals)? If yes, what was the intended population?

B. *Sampling*: Did the authors use a sampling plan that ensured the representativeness of their sample for the intended population? If yes, how did they do this (e.g., random sampling from population, inclusion/exclusion criteria)?

C. *Attrition*: Did the authors implicitly or explicitly provide assurance that there were no problems related to attrition? If yes, how was this done?

### *II. Settings*

A. *Setting*: Did the authors state the intended setting (e.g., restaurant, cafeteria, grocery store, home)? If yes, what was the intended setting?

B. *Representativeness*: Did the authors take steps to ensure that the study setting was representative of the intended setting? If yes, how was this done?

## *III. Outcomes*

A. *Measures*: Did the authors measure the most appropriate dependent variables (e.g., weight of participants) rather than proxies (e.g., amount of food consumed)? Whether yes or no, what dependent variables were measured?

B. *Reporting*: Did the authors report findings in ways that address important goals (e.g., reaching or maintaining a healthy weight, BMI, or obesity status)? If yes, what was reported?

C. *Compensation*: Did the authors consider mechanisms that might counteract apparent effects (e.g., behavioral effects such as adaptation over time or eating less in other settings to offset greater consumption during the study)? If yes, what kinds of compensation were considered?

D. *Adverse consequences*: Did the authors address potential harms (e.g., inconvenience, negative emotions, reduced quality of life)? If yes, what potential harms were addressed?

E. *Moderators*: Did the authors report any analyses of moderator effects to assess the robustness versus specificity of effects (e.g., healthy vs. unhealthy participants, age differences, cultural differences)? If yes, what moderators were tested?

F. *Sensitivity*: Did the authors perform any sensitivity analyses to assess dose-response effects, threshold levels, or diminishing returns? If yes, how was this done?

G. *Costs*: Did the authors discuss costs (e.g., money, time) of putting proposed interventions into practice? If yes, how was this done?

## *IV. Time frames*

A. *Duration*: Did the authors study a phenomenon over a sufficient period of time to avoid having to extrapolate from their findings? (In other words, were there repeated exposures, or was this a one-off observation?) Whether yes or no, what was the duration of the study?

B. *Long-term effects*: Did the authors report data on longer-term effects following the intervention? If yes, how long was the follow-up period?

C. *Sustainability*: Did the authors consider the sustainability of the intervention after the formal evaluation? If yes, was this done through data analysis or discussion?

### Action Editor

Laura A. King served as action editor for this article.

### Author Contributions

J. Ruscio and A. Zabel developed the study concept. All authors contributed to the sampling plan and coding scheme. Data collection was performed by A. B. Edelblum, L. Hatch, C. M. Loyka, and B. Wetreich, and these four individuals also calculated the coding agreement. C. M. Loyka compiled, organized, checked, and made necessary corrections and updates to the master data files. J. Ruscio supervised data analysis, recorded citation counts, and performed the correlational analyses. C. M. Loyka drafted the manuscript, and all authors provided comments and suggestions for revisions, which were performed by J. Ruscio. After C. M. Loyka and J. Ruscio, authors are listed alphabetically because they made approximately equal contributions. All of the authors approved the final manuscript for submission.

### ORCID iD

Caitlin M. Loyka https://orcid.org/0000-0001-7528-1116

### Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

### Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/1745691619876279

### References

Amir, O., Ariely, D., Cooke, A., Dunning, D., Epley, N., Gneezy, U., . . . Silva, J. (2005). Psychology, behavioral economics, and public policy. *Marketing Letters*, *16*, 443–454.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Casazza, K., Fontaine, K. R., Astrup, A., Birch, L. L., Brown, A. W., Bohan Brown, M. M., . . . Allison, D. B. (2013). Myths, presumptions, and facts about obesity. *New England Journal of Medicine*, *368*, 446–454. doi:10.1056/NEJMsa1208051

Clarke, G. N. (1995). Improving the transition from basic efficacy research to effectiveness studies: Methodological issues and procedures. *Journal of Counseling and Clinical Psychology*, *63*, 718–725. doi:10.1037/0022-006X.63.5.718

Coolican, H., & Flanagan, C. (2005). The study lacks ecological validity. *Psychology Review*, *11*(3), 24–26.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine*, *15*, 451–474. doi:10.1016/0091-7435(86)90024-1

Glasgow, R. E., Green, L. W., Klesges, L. M., Abrams, D. B., Fisher, E. B., Goldstein, M. G., & Tracy Orleans, C. (2006).

External validity: We need to do more. *Annals of Behavioral Medicine*, *31*, 105–108. doi:10.1207/s15324796abm3102_1

Green, L. W., & Glasgow, R. E. (2006). Evaluating the relevance, generalization, and applicability of research. *Evaluation and the Health Professions*, *29*, 126–153. doi:10.1177/0163278705284445

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 429–446). Hillsdale, NJ: Erlbaum.

Jüni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. *The British Medical Journal*, *323*(7303), 42–46. doi:10.1136/bmj.323.7303.42

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus, and Giroux.

Kerner, J., Rimer, B., & Emmons, K. (2005). Introduction to the special section on dissemination: Dissemination research and research dissemination: How can we close the gap? *Health Psychology*, *24*, 443–446. doi:10.1037/0278-6133.24.5.443

Klesges, L. M., Dzewaltowski, D. A., & Glasgow, R. E. (2008). Review of external validity reporting in childhood obesity prevention research. *American Journal of Preventive Medicine*, *34*, 216–223. doi:10.1016/j.amepre.2007.11.019

Prohaska, T. R., & Etkin, C. D. (2010). External validity and translation from research to implementation. *Generations*, *34*, 59–65.

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. doi:10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128. doi:10.1177/1745691617708630

Steckler, A., & McLeroy, K. R. (2009). The importance of external validity. *American Journal of Public Health*, *98*, 9–10. doi:10.2105/AJPH.2007.126847

Teachman, B. A., Norton, M. I., & Spellman, B. A. (2015). Memos to the President from a "Council of Psychological Science Advisers." *Perspectives on Psychological Science*, *10*, 697–700. doi:10.1177/1745691615605829

Wansink, B. (2006). *Mindless eating: Why we eat more than we think*. New York, NY: Bantam Books.

Wansink, B. (2014). *Slim by design: Mindless eating solutions for everyday life*. New York, NY: HarperCollins.

Wansink, B., & Chandon, P. (2014). Slim by design: Redirecting the accidental drivers of mindless overeating. *Journal of Consumer Psychology*, *24*, 413–431. doi:10.1016/j.jcps.2014.03.006

Wansink, B., Painter, J. E., & North, J. (2005). Bottomless bowls: why visual cues of portion size may influence intake. *Obesity Research*, *13*, 93–100.