RESEARCH ARTICLE

Gaming Anthropology: The Problem of External Validity and the Challenge of Interpreting Experimental Games

Nicole Naar

American Anthropologist Vol. 122, No. 4 December 2020

ABSTRACT Experimental economic games are an increasingly common component of the anthropological tool kit. Yet their external validity continues to be a point of debate and active empirical investigation within economics and anthropology. I review and reorganize central concepts within the experimental economic game literature on external validity and find that—consistent with anthropological assumptions of cultural variability—game results are not reliably generalizable across different participants or contexts. However, whether or not game behavior parallels real-world behavior within the same participants or contexts remains an open question. Methodological diversity is a strength of anthropology as a discipline, and therefore anthropologists are well poised to design more effective tests of parallelism in the future. In the meantime, anthropologists borrowing experimental methods from economics should treat the relationship between behavior inside and outside of games as an open empirical question. They should also carefully consider whether the method is consistent with their theoretical assumptions and research goals. [*experimental economic games, external validity, generalizability, parallelism, anthropological methods*]

RESUMEN Juegos económicos experimentales son un componente cada vez más común del equipo de herramientas antropológicas. Sin embargo, su validez externa continúa siendo un punto de debate e investigación empírica activa dentro de la economía y la antropología. Reviso y reorganizo conceptos centrales dentro de la literatura de juegos económicos experimentales sobre la validez externa y encuentro que –consistente con asunciones antropológicas de variabilidad cultural– los resultados de los juegos no son generalizables de forma confiable a través de diferentes participantes o contextos. Sin embargo, si el comportamiento del juego paralela o no el comportamiento del mundo real dentro de los mismos participantes o contextos permanece como una pregunta abierta. Diversidad metodológica es una fortaleza de la antropología como una disciplina, por lo tanto, antropólogos están bien posicionados para diseñar pruebas más efectivas de paralelismo en el futuro. Entre tanto, antropólogos prestando métodos experimentales de la economía podrían tratar la relación entre el comportamiento dentro y fuera de los juegos como una pregunta empírica abierta. Deberían también considerar cuidadosamente si el método es consistente con sus asunciones teóricas y metas de investigación. [*juegos económicos experimentales, validez externa, capacidad de ser generalizable, paralelismo, métodos antropológicos*]

1. INTRODUCTION

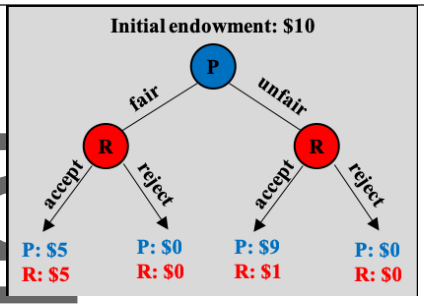
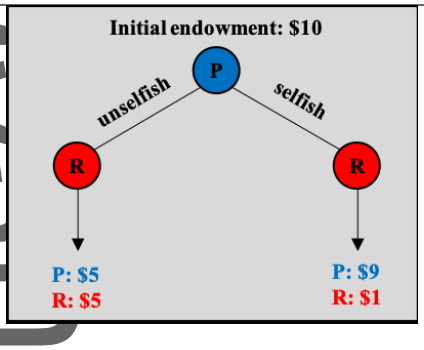
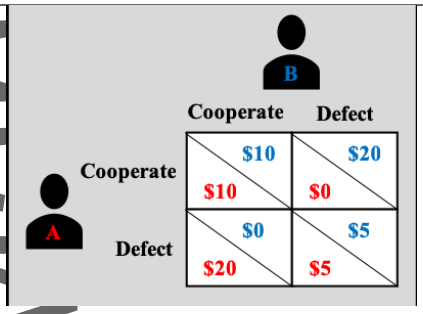
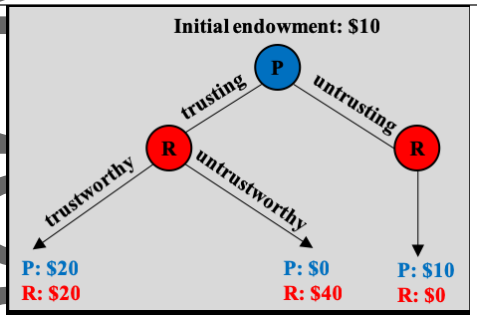
Since experimental economic games (EEGs) first made their way from the economics laboratory to the anthropological fieldsite nearly two decades ago (Henrich 2000), they have become an increasingly popular and flexible methodological tool for evolutionary and economic anthropologists. Originally rooted in the assumptions of the rational actor model in economics, EEGs isolate and manipulate specific constraints to study their influence on decision-making and behavior. Anthropologists have been particularly interested in economic experiments designed to measure social

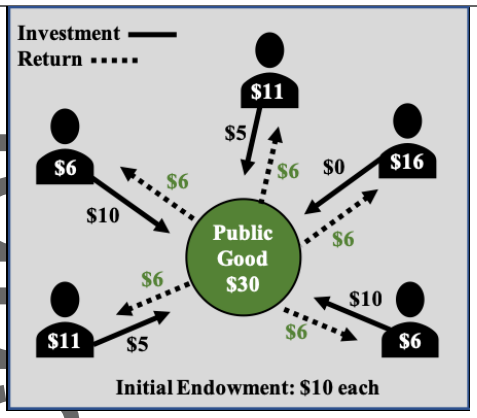
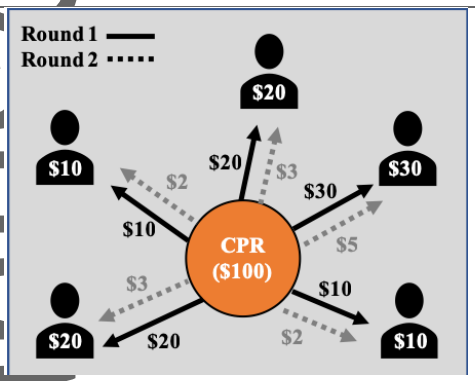
preferences (Table 1), and their elaborations on classical EEGs have expanded both the pool of game participants (e.g., Brosnan 2013; Henrich et al. 2005) and the range of hypotheses considered (e.g., Bauer et al. 2014; Gervais 2017; Pisor and Gurven 2018; Purzycki et al. 2016). Compared to ethnographic observations of behavior or the collection of survey data, experimental approaches provide multiple advantages: (1) they allow researchers to systematically compare observed behavior to the analytically rigorous predictions of game theory; (2) the controlled and replicable environment provides methodological consistency; (3) data can be collected quickly by a small number of researchers from a large sample of participants; (4) they can be used to invoke behaviors that are otherwise difficult to observe in real time or in naturalistic settings; and (5) their widespread use facilitates cross-disciplinary communication and collaboration (Ensminger 2002; Lesorogol 2017).

[TABLE 1 ABOUT HERE]

Table 1. A description of EEGs frequently administered in anthropological research settings and the social preferences they were designed to reveal. For alternative visualizations and descriptions, see Luo and Yu (2015) and Levitt and List (2007a), respectively.

Name	Visual Description	Verbal Description	Social Preferences[1]

Ultimatum Game (UG)		The proposer (P) decides how to split a fixed sum of money with the responder (R). If R accepts, the money is allocated as proposed. If R rejects, neither player receives any money.	P: fairness, inequity aversion R: fairness, inequity aversion, negative reciprocity
Dictator Game (DG)		A variation on the UG. P decides how to split a fixed sum of money between P and R, and the money is allocated as proposed. Many allocation games used by anthropologists are elaborations of the DG.	P: altruism, fairness, inequity aversion
Prisoner's Dilemma (PD)		Two players simultaneously decide to cooperate or defect. The optimal choice for each player is in conflict with the group-optimal choice.	altruism, (conditional)[2] cooperation
Trust Game (TG)		A sequential PD. P decides how to split an initial endowment between P and R. All money sent to R increases by a factor (four, in this example). R then decides how to split the increased sum of money between R and P.	P: trust, positive reciprocity R: trustworthiness, positive reciprocity

Public Good Game (PGG)[3]		An n-person PD usually played in multiple rounds. Each player is given an initial endowment to either invest or keep. Invested funds increase by a factor and then are evenly redistributed to all players.	altruism, (conditional)[2] cooperation
Common-Pool Resource Game (CPR)[3]		An n-person game resembling a "non-linear public bad"[4] that is also played in multiple rounds. Players decide how much to harvest from a rivalrous, nonexcludable resource.	altruism, (conditional)[2] cooperation

[1]P refers to how proposer game behavior is interpreted, and R refers to how responder game behavior is interpreted, in terms of social preferences.

[2]The assumption is that one-shot games measure cooperation, while sequential games with multiple rounds also measure conditional cooperation.

[3]Final payoff for each player is indicated in white for the PGG and CPR game. In the CPR game example, the resource was depleted to $10 after one round. Harvests in round 2 exceeded this amount, resulting in no additional payoffs.

[4]Cardenas and Carpenter (2008, 313).

In spite of their strengths, however, the narrow context and simplifying assumptions that lend rigor to EEGs may limit their relevance to real-life situations, prompting concerns about their external validity. The external validity of EEGs has been an ongoing point of debate both within economics and anthropology, motivating a small but growing literature. Informing these debates are deeper questions: Does the external validity of EEGs matter? And if so, when does it matter? These are crucial questions, especially for anthropologists, because many collect data using ethnographic methods such as participant observation, surveys, or interviews. If EEG results are not meaningfully related to behavior observed in nonexperimental contexts, what is their value for anthropologists?

Within economics, ensuring internal validity—or the ability to draw causal conclusions from EEGs—is paramount, especially when the goal is testing theory (Schram 2005). EEGs were, in fact, designed to be simplified and abstract to strip away the contextual richness that complicates inferring causal relationships from observational studies. By providing temporal clarity, allowing for random assignment to treatment groups, and increasing researcher control over context and setting, experiments minimize interference from potential confounds and maximize replicability (Roe and Just 2009). For example, if a researcher wants to test theories of human altruism, they might use a dictator game with the following design features to maximize internal validity: randomly selected participants, computer-lab setting, tokens as currency, complete anonymity, no visual contact or communication between players, and written instructions using intentionally general or neutral language.

External validity, in contrast, refers to the ability to apply conclusions from EEGs to other situations. Depending on the situation, the specific details may change. If applying a dictator game to the context of local school donations, for example, a game that maximizes external validity might instead be called "the donation game" and be administered by a teacher in a classroom setting (i.e., visual contact) to parents of school-age children with multiple currencies (i.e., time and money) at their disposal. Rather than being "the Achilles heel of all laboratory experimentation," as Loewenstein claimed (1999, F33), external validity is more often viewed as trading off with internal validity (Roe and Just 2009). Thus, increasing external validity may pose an unwarranted threat to internal validity unless it is relevant to the specific research question (Druckman and Kam 2011; Gächter 2010) or unless the goal is making policy recommendations (Camerer 2015; Schram 2005). Even though most experiments in economics are theory- rather than policy-driven, "experimental economists are very active in advising policy makers" (Schram 2005, 233). Self-awareness of this fact has prompted greater attention to the challenge of external validity among economists.

Many anthropologists (and others) have suggested, on the other hand, that the apparent lack of external validity of EEGs threatens their internal validity—regardless of the goal—because it

suggests the games may not be measuring what we think they are measuring. If the social preferences displayed in EEGs do not reflect behavior in everyday life, then do EEGs truly measure stable social preferences? Though anthropologists may borrow EEGs as a methodological tool, the associated disciplinary assumptions transfer less easily. Social anthropologists have questioned the assumption that a single game can measure abstract concepts as cross-culturally variable and contextually flexible as fairness, generosity, and cooperation (Chibnik 2005; Jackson 2012). They are also skeptical of the efficacy of experimental "controls" aimed at eliminating the "noise" generated by multiple competing norms, motivations, and identities all up for "individual interpretation and creative manipulation" (Jackson 2012, 4). Similarly, evolutionary anthropologists and psychologists have called attention to adaptive motivators and cues existing alongside the short-term incentive structure of EEGs that may also affect intuited payoffs and, by extension, game behavior (Hagen and Hammerstein 2006; E. Smith 2005). For example, if reputational concern is an adaptive feature of human psychology that helps individuals achieve long-term benefits in spite of short-term costs, how should we interpret altruistic behavior in EEGs? How do we know which constraints game participants are responding to—the short-term constraints of the game (e.g., anonymity, one-shot) or the longer-term constraints of everyday life (e.g., no anonymity, repeated interaction)?

The conceptual challenges raised by anthropologists suggest that the link between theory and evidence may be more complicated than assumed for EEGs. If game behavior does not resemble behavior in the context of daily life, how can we assume it reflects a stable social preference or an internalized social norm? In other words, in the absence of external validity, how do we interpret game results? Such questions are particularly pressing for applied anthropological research and for questions with practical implications. For example, among scholars of the commons—many of whom are anthropologists—working toward external validity has become essential for making sense of EEG results (Anderies et al. 2011).

Here I review, reorganize, and summarize the literature on the external validity of EEGs, with a focus on the theoretical and methodological implications of their use by anthropologists. Although external validity has been conceptualized and organized in various ways (see, for example, Fréchette [2015] and footnote 2 in Levitt and List [2007b]), I propose distinguishing between two main aspects of external validity: generalizability and parallelism (Figure 1). Generalizability refers to the extent to which the observed effects apply to other populations, subject pools, settings, and contexts (Campbell and Stanley 1963). Or, more directly, do EEG results hold *across different* subjects or contexts when the *same* game is played? While tests of generalizability unpack universalizing theoretical assumptions, concerns about parallelism raise important methodological questions, such as whether behavior in the game reflects behavior in daily life under similar conditions (Shapley 1964; cited in Levitt and List 2007b). In other words, *within the same* subject or context, are EEG results consistent with *different* measures of similar behaviors? The literature as a whole suggests that EEGs do not reliably generalize, but tests of parallelism reveal no clear consensus.

Conceptualizing external validity in this way helps anthropologists move beyond general concerns about external validity to specific questions about which aspects of external validity are relevant to their research questions and when external validity should be expected given their theoretical assumptions. The distinction between generalizability and parallelism also draws a clear line between previous research by anthropologists that has challenged core assumptions in economics and issues that should inform future research on external validity. Addressing questions of parallelism requires the methodological pluralism familiar to many anthropologists, providing yet another opportunity for anthropology to influence other disciplines. However, before designing new tests of parallelism, we must first clarify our own theories of institutions and evaluate how well (or how poorly) they are reflected in the EEGs we select. Finally, researchers who conduct long-term fieldwork and cultivate close relationships with informants should also consider the methodological implications of using experimental methods that may have long-term impacts.

[FIGURE 1 ABOUT HERE]

My review of the external validity literature is thorough, but targeted. I focus on social preference games designed to reveal preferences about cooperation, sharing, and fairness, and thus exclude the large experimental literatures on market institutions and voting in economics and political science, respectively. Within the literature on social preference games, I pay particular attention to explicit tests of external validity. I used Google Scholar to search the literature for studies that specifically examine one or more of the following dimensions of generalizability: population, subject pool, setting, and contextual frame. Many EEG analyses include some external indicator—even if they do not explicitly test for parallelism. To narrow this part of the review, I emphasize games designed to better understand decisions about resource use (i.e., PGG and CPR games), where issues of ecological validity and policy application are particularly salient. I also include studies using other social preference games (e.g., DG, UG, TG) if they are contributions by anthropologists or explicit tests of parallelism. Finally, any reviewed generalizability studies that also speak to issues of parallelism are included in the parallelism review.

2. GENERALIZABILITY

Given that EEGs were designed under disciplinary assumptions of universal preferences and psychology, the behavior observed in them was most often interpreted as a response to the game's incentives rather than outside factors. However, for decades, the majority of EEG research involved abstract games played with student subjects from WEIRD (Western, educated, industrialized, rich, democratic) populations in laboratory settings (Henrich, Heine, and Norenzayan 2010). This left many

wondering whether similar patterns of behavior in EEGs reflected universal preferences or a lack of variation in population, subject pool, setting, and/or contextual frame. Although these terms often overlap within anthropology, in the EEG literature, *population* refers to particular countries or cultural groups, *subject pool* refers to specific subgroups within populations (i.e., women, children, university students), and *setting* refers to the location of game administration (i.e., computer lab, classroom, field). Concerns about the generalizability of EEGs across these dimensions are not new (e.g., Campbell and Stanley 1963), but there was a renewed interest in empirically testing the generalizability of EEGs once anthropologists began running games in the field with their informants.

2.1. Across Populations

To date, many researchers in both economics and anthropology have tested whether the "typical findings" of EEGs hold across different populations (i.e., countries, cultural groups, communities) (Figure 2 and Table S1). The widely cited study of EEGs played in fifteen small-scale societies (Henrich et al. 2005) confirmed on a larger scale the conclusions of Henrich's (2000) initial study comparing Los Angeles students and Peruvian forager-horticulturalists: the prosocial preferences and game behavior of WEIRD populations do not generalize to other economies or cultures. Cross-country comparisons have yielded similar conclusions. A few have found similar patterns of EEG behavior across different Western (e.g., Franzen and Pointner 2013) and non-Western populations (e.g., Cameron 1999). However, taken as a whole, the literature suggests that EEG results do not generalize across populations within or between different cultural and geographic regions. Even within a single country, subpopulations based on race/ethnicity, village, livelihood, and political institution exhibit significant differences in patterns of gameplay (Table S1).

[FIGURE 2 ABOUT HERE]

Given the various effects of culture, economy, policy, and ecology, it is still unclear what drives this variation across populations. In a meta-analysis of cultural differences in ultimatum game play, Oosterbeek, Sloof, and van de Kuilen (2004) detected regional-level differences but could not explain those differences based on specific cultural traits. Similarly, Paciotti and Hadley (2003) found differences in EEG behavior between ethnic groups residing in the same villages, while Prediger, Vollan, and Frolich (2011) found the reverse; distinct political and ecological conditions in different countries generate differences in EEG behavior within the same ethnic group. Other proposed sources of population differences include group-specific cultural norms (Henrich et al. 2005), social expectations (Gurven, Zanolini, and Schniter 2008), political history, and local ecology (Prediger, Vollan, and Frolich 2011). Regardless of the specific source of population differences, however, the empirical record as a whole strongly suggests that the results of EEGs reflect population-level variation rather than universal patterns of behavior.

2.2. Across Subject Pools

Another area of concern for external validity is the generalizability of game results across different subject pools (Figure 2 and Table S2). Though many aspects of subject pool bias may influence results (e.g., gender, see Croson and Gneezy 2009), most empirical tests have focused on the representativeness of student and/or volunteer subjects that constitute the "narrow data base" (Sears 1986) of most laboratory experiments. Nonstudent adult subjects are often more difficult to recruit and retain given their physical and educational distance from the university lab. Therefore, relying on student subject pools is considered both convenient and reliable (Hooghe et al. 2010). But researchers have long cautioned that student subject pools systematically differ from more-representative adult samples (Sears 1986; cf. Druckman and Kam 2011). In the majority of empirical studies explicitly

comparing students and nonstudents, students as a whole are less prosocial, less trusting, less generous, and/or less cooperative than other adults (see Table S2).

When students are compared to nonrepresentative groups of adults, there is less consensus surrounding generalizability. In a literature review of student subjects versus professional subjects, Fréchette (2015) concludes that the overall qualitative results are similar. The few studies that directly compare students to professionals support the general finding that students are less prosocial (Carpenter and Seki 2011; Fehr and List 2004; Stoop, Noussair, and van Soest 2012). But in one study (Fiedler and Haruvy 2009), students were more trusting and trustworthy when playing an online virtual-world game than experienced players recruited from the internet.

Significant differences between subgroups of students have also been found. For example, Carpenter, Burks, and Verhoogen (2005) found that nontraditional students at a community college made a larger number of generous offers in the ultimatum game than traditional university students. Another small but growing literature focusing on the representativeness of economics majors—the most convenient of convenience samples—suggests that economics students are less prosocial than noneconomics students (for a recent review, see Gerlach 2017). In fact, Frigau, Medda, and Pelligra (2019) found that much of the difference between students and nonstudents in dictator game offers was driven by economics majors.

Selection bias might also generate differences between subject pools and produce results that fail to generalize. Slonim et al. (2013) note that few studies report participation rates, and those that have reported participation rates as low as 2 percent. This raises an important question: Are there systematic differences between those who choose to participate in EEGs and those who decline? The empirical record is inconsistent. Although some authors have found significant differences between the attributes, preferences, and behavior of volunteers and pseudo-volunteers (i.e., subjects who participate in experiments as part of a college course), others have found no such differences (see

Table S2). Overall, there is general consensus in the literature that the results of EEGs conducted with student subjects do not generalize to other subject pools, but the effects of selection bias on generalizability remain ambiguous.

2.3. Across Settings

Generalizability across different EEG settings is also frequently mentioned as an important element of external validity (Figure 2 and Table S3). It is hypothesized that the heightened scrutiny of the laboratory setting increases cooperation relative to the field (Levitt and List 2007a). However, Stoop, Noussair, and van Soest (2012) found that recreational fisherfolk behaved more cooperatively in artifactual field experiments than in laboratory experiments, and Beramendi, Duch, and Matsuo (2016) found no significant difference between student game behavior in laboratory and online settings. Other studies have sought to empirically test the effect of setting on EEG behavior, but the influence of setting is often unclear because different subject pools play the games in different settings or because abstract EEGs are compared to context-rich natural field experiments. Englmaier and Gebhardt (2016) compared laboratory experiments and natural field experiments for three games, each with a different incentive structures (group, individual, and none). They found positive correlations between lab and field behavior, but only in the group-incentives games. Conversely, Galizzi and Navarro-Martinez (2019) found no correlation between behavior in the lab and behavior in various NFEs. In a study looking only at behavior in a natural field experiment resembling the dictator game, Winking and Mizer (2013) reported that no subject offered any part of their windfall endowment to a stranger. This is in stark contrast to results from laboratory versions of the dictator game in the broader EEG literature, suggesting that people may indeed behave more pro-socially in laboratory settings.

Given these inconsistent results, more research that effectively controls for subject pool, contextual frame, and methodological effects is needed before we can reach a broader consensus on the generalizability of EEGs across different settings. When student subjects in the laboratory and adult subjects in the field behave differently, is the difference one between subject pools (e.g., students/adults), settings (e.g., lab/field), or contexts/social frames (e.g., school/work)? Furthermore, if participants do not realize they are taking part in an experiment—as is the case with NFEs—should they be compared to participants in laboratory experiments as evidence of generalizability across settings, or should these results be considered distinct behavioral measurements that instead speak to questions of parallelism? Indeed, the only difference between testing generalizability across settings and testing parallelism is a shift from between-subjects to within-subjects design.

2.4. Across Contextual Frames

One final aspect of generalizability that is often perceived as a threat to the external validity of EEGs is the sensitivity of game behavior to subtle changes in the contextual framing of the game (Levitt and List 2007a). Though other aspects of context (broadly defined) have been studied—including different incentives (e.g., Englmaier and Gebhardt 2016), different parameters and group sizes (e.g., Goeschl et al. 2015), and valence framing effects—here I focus on contextual framing effects (Figure 2 and Table S4). Hagen and Hammerstein (2006, 345) define a frame as "a knowledge structure or conceptual abstraction used to interpret a complex reality or experience, and guide behavior." Even though classical EEGs are intentionally abstract and devoid of context, participants may provide their own frame by invoking relational cues, social and cultural norms, decision-making heuristics, and past experiences. Under such conditions, the experimenter has little control over which frame(s) players are using to make sense of the game or whether or not subjects are relying on the same frames. For example, rather than simply playing an abstract public goods game, it quickly became clear during

postgame discussions that many Orma players were instead thinking of the *harambee*—a local fundraising institution (Henrich et al. 2005). Alternatively, an experimenter might deliberately introduce a contextual framing treatment by comparing unframed and framed versions of the same game. A meat-sharing game and unframed dictator game, for example, might have the same incentive structure yet produce very different results because of contextual cues (Lesorogol 2007).

Precisely how framing effects influence game behavior is an issue of theoretical debate (Gerlach and Jaeger 2016), but several empirical studies of contextual framing effects suggest that changes in game frame induce changes in game behavior. In prisoner's dilemma and public goods games, both of which measure cooperation, economic and noneconomic frames produce different results (e.g., Ellingsen et al. 2012), and even shifts as subtle as changing the pronouns used in game instructions ("I" versus "we") can affect game behavior (Cookson 2000). Context framing effects have also been found in ultimatum games, dictator games, and trust games (Table S4). Anthropologists have found significant differences between behavior in abstract games and games contextualized in specific cultural institutions. In most cases, the selected cultural institutions were familiar to game participants (Cronk 2007; Gerkey 2013; Lesorogol 2007), but Cronk and Wasielewski (2008) found that even unfamiliar cultural frames produce framing effects. In summary, as a whole the empirical record suggests that EEGs do not reliably generalize across different contextual frames.

3. PARALLELISM

A separate concern about external validity focuses on questions of parallelism. Formally defined as the transferability of propositions tested in the lab to "nonlaboratory microeconomies where similar ceteris paribus conditions hold" (V. Smith 1982, 936), tests of parallelism compare EEG results to behaviors or attributes in naturalistic contexts. For example, a researcher might wonder if a player's

altruistic allocations in a dictator game reflect behavior in other situations. Numerous studies have been designed to empirically address concerns about parallelism, but with no consistent criteria for designing tests of parallelism, the results and their interpretation have been inconsistent. Part of the issue stems from the abstract nature of classical EEGs, which at best only resemble concrete real-world scenarios in the underlying game structure. This makes it easy to rationalize any real-world behavior as somehow relating to behavior observed in the EEG (Galizzi and Navarro-Martinez 2019). For this reason, I propose two general prerequisites for establishing external validity in terms of parallelism: context parallelism and indicator parallelism.

3.1. Context Parallelism

Given the well-documented influence of framing effects on EEG behavior (see above), researchers have paid increasing attention to context parallelism, or how similar the decision-making context of the EEG is to the real-world context of interest (Figure 3a and Table S5). In other words, does the EEG have what some economists call "ecological validity"?[1] One way to enhance context parallelism is to choose an explicit frame. For example, Benz and Meier (2008) frame a dictator game as donations to social funds and find that student donations to those same social funds in real life are weakly correlated with laboratory results. Similarly, Gelcich and colleagues (2013) found that fishing-union success was related to group performance in a common-pool resource game framed as a fishery. However, relatively more-parallel contexts do not always produce more parallel behavioral results. In

[1] Some economists view ecological validity as one component of external validity (Levitt and List 2007b), while others describe external validity and ecological validity as distinct concepts (see Fréchette 2015; Roe and Just 2009). The reorganization suggested in Figure 1 is a compromise between these two positions.

other studies where fisherfolk played framed EEGs, external indicators of behavior were not associated with game behavior (Javaid et al. 2016; Stoop, Noussair, and van Soest 2012).

In addition to framing the EEG, another way to increase context parallelism is to ensure the underlying game form and parameters reflect the real-world situation of interest. Recently developed common-pool-resource experiments for forests, irrigation systems, fisheries, and grazing lands enhance ecological validity by incorporating specific resource dynamics, including nonlinearity, spatial variability, path dependence, and/or asymmetrical access (Cárdenas, Janssen, and Bousquet 2013). Institutional elaborations reflecting alternative management strategies can also be incorporated into the games to reflect the interplay between social and ecological systems. Very few studies have explored whether these efforts to enrich the context parallelism of EEGs produces greater parallelism between game and real-life behaviors. Prediger, Vollan, and Frolich (2011) found that political and ecological history reliably predicts group performance in a grazing game, and Handberg and Angelsen (2015) found that individual extraction in a forestry game was positively correlated with relative forest use. However, a recent analysis comparing behavior in a framed fishing game to survey responses and ethnographic observations found little evidence of parallelism (Naar et al., submitted).

[FIGURE 3 ABOUT HERE]

3.2. Indicator Parallelism

Apart from ecological validity or context parallelism of the game itself, the selection of external indicators also raises important issues regarding parallelism. As Torres-Guevara and Schlüter (2016) point out in their review of previous studies of external validity in CPR systems, relationships between game and real life have been strongest when the selected behavioral measures are themselves strongly linked. In studies using unframed EEGs, the external indicator can be virtually any prosocial

behavior. This diversity of external indicators mirrors the diversity of results (see Figure 3b and Table S6), with as many positive as negative and ambiguous conclusions regarding external validity. In a recent systematic review and meta-analysis of parallelism between laboratory and field measures of social preferences, Galizzi and Navarro-Martinez (2019) report that less than 40 percent of studies find statistical evidence of external validity. But even framed games with high context parallelism may be lacking in indicator parallelism if game behavior is only indirectly related to behavior in daily life. For example, a CPR game framed as a fishery may have high context parallelism, but fishing effort has greater indicator parallelism than fishing-cooperative membership.

Another related area of concern is the level at which the external indicator operates (Torres-Guevara and Schlüter 2016). Though many studies compare individual behaviors, others relate individual game behaviors to real-life group affiliations (e.g., village or household membership), and some compare the behavior of groups. Interestingly, relatively more group-level analyses report evidence of external validity compared to individual-level analyses (see Figure 3b and Table S6). In one puzzling case (Bouma, Bulte, and van Soest 2008), there was no evidence for parallelism at the individual level, but aggregate game results at the village level predicted individual contributions to water and soil contributions in real life. Perhaps individual-level behavior is inconsistent across measures, but group-level differences in norms can affect both gameplay and real-world behavior. This suggests that we should pay close attention to the source and level of variation in behavior. In summary, the ambiguous evidence of parallelism between behaviors inside and outside of EEGs could therefore stem from any combination of differences in research design, including varying degrees of parallelism between contexts, behavioral tasks, and levels of analysis.

4. IMPLICATIONS FOR ANTHROPOLOGICAL RESEARCH

Viewed from an anthropological perspective, the EEG literature on external validity raises important theoretical and methodological issues that warrant further consideration.

4.1 Theoretical Issues

The overall finding that EEG results do not reliably generalize is perhaps both unsurprising and unproblematic for many anthropologists who are disciplinarily inclined to expect variability. The same is true for the many economists who are no longer committed to the rational actor model (e.g., Sen 1977) and instead view these frequent departures from rationality in EEGs as opportunities to better understand human behavior (Levitt and List 2008; Thaler 2000). However, different forms of generalizability pose different opportunities and challenges, especially from an anthropological perspective. When is a lack of generalizability a problem for anthropologists, and when is it an asset? Poor generalizability across populations and contexts is likely a key strength of EEGs for anthropologists because it validates our discipline's core assumptions of cultural and contextual variation. However, unless an anthropologist has reason to expect differences between subject pools (e.g., gender, age, class), poor generalizability in this area might be problematic. On the one hand, evolutionary anthropologists may expect certain institutions or cultural norms (e.g., fairness) to be uniformly expressed across different subject pools within a single population (Henrich et al. 2005). On the other hand, sociocultural anthropologists might view systematic differences between subject pools as a logical extension of contextual or institutional variation (Jackson 2012). Similarly, biological anthropologists could predict differences between subject pools based on life history theory (e.g., Cassar, Wordofa, and Zhang 2016). Comfort with a lack of generalizability in any particular area ultimately depends on one's theoretical commitments and research questions.

The inconsistent parallelism between behaviors inside and outside of EEGs also raises important theoretical issues. When should parallelism matter in anthropological research, and what are the implications if/when EEG results bear little resemblance to ethnographic observations? The

answer again depends, in part, on the research question(s). In some cases, parallelism is a completely irrelevant concern. Indeed, one perceived benefit of EEGs is that they allow researchers to observe preferences while removing or manipulating the constraints of daily life—counterfactuals that are unobservable using traditional ethnographic methods (Pisor et al. 2020). For example, the anonymity of EEGs allows anthropologists to study the influence of kinship (e.g., Macfarlan and Quinlan 2008), friendship (Rucas et al. 2010), and other institutions on economic decision-making in the absence of social pressures—a rare and difficult-to-observe situation in many small-scale societies. However, other lines of anthropological inquiry either implicitly or explicitly rely on the assumption that EEG behavior relates to observed behavior. For example, studies that aim to measure preferences would expect game behavior to reflect observed behavior, assuming the constraints of the game also sufficiently parallel real life. Games designed to re-create one or more of the constraints imposed by local institutions also assume or strive for some degree of parallelism.

Whether or not researchers should expect external validity and be concerned by its absence also depends on a much deeper issue: the underlying theory of institutions informing their assumptions and interpretations. EEGs were designed to test the relationship between beliefs, preferences, and constraints (Gintis 2006); and although the language of institutions infuses the EEG literature, economists and anthropologists may have implicitly different ideas about where institutions fit into the preferences, beliefs, and constraints approach. In an insightful review of experiments in ecological economics, Rommel (2015) identifies multiple conceptualizations of institutions and suggests many applications of EEGs fail to clearly and consistently link theory and method. His distinction between structural and agent-based understandings of institutions is particularly relevant. “Institutions can be viewed as structures exogenous to the agent” (98), which Rommel links to classical economic experiments on alternative institutional arrangements. In other words, changing the institution (e.g., payment for ecological services, community forest management, and command and control) changes the incentive structure, thereby changing behavior (e.g., forest conservation)

(Handberg and Angelsen 2015). Alternatively, institutions can be theorized "as cognitive media embedded solely in the agent" (Rommel 2015, 98), making them more subjective and difficult to experimentally manipulate. This perspective lends itself to the cross-cultural comparisons (e.g., Henrich et al. 2005; Henrich et al. 2006) and studies of framing effects (Cronk 2007; Cronk and Wasielewski 2008; Lesorogol 2007) that many anthropologists have spearheaded.

Rommel does not venture into the external validity debate, but different theories of institutions imply different expectations of external validity when they are methodologically linked to EEGs. If institutions are external structures (i.e., constraints)—as experimental economists are likely to assume—then removing or altering constraints can reveal underlying preferences and enable game participants to exhibit behavior that would be impossible or irrational in real life. At the outset, such a theoretical approach negates a concern with parallelism; but if universal preferences are assumed, generalizability does become a concern. If, however, institutions are internalized in agents (i.e., norms, preferences, and/or beliefs)—as anthropologists are more likely to assume—then behavior in games reflects the strength of shared norms, and similar patterns of behavior can be observed within communities (for different individuals) and across contexts (for the same individuals). This theoretical approach instead takes generalizability as irrelevant but expects to find evidence of parallelism, especially if the frame is crafted to cue shared norms and the incentives are designed to mimic real-world constraints. In the absence of parallelism, one is left wondering if EEGs are capable of controlling and measuring their intended targets. Disagreements about the relative embeddedness of norms and institutions also implicitly inform similar debates about the role of anonymity, stakes, and currencies in EEGs (Chibnik 2005; Henrich et al. 2005; Jackson 2012; E. Smith 2005; Sullivan and Lyle 2005).

Evolutionary theories of institutions are one possible way forward. By emphasizing the interdependence of institutional structures and agents' internalization of norms and preferences (Rommel 2015), evolutionary perspectives in institutional economics define institutions as

"simultaneously both objective structures 'out there' and subjective springs of human agency 'in the human head'" (Hodgson 2006, 8). This resonates with the ideas of practice theorists (e.g., Bourdieu 1977; Giddens 1986; Ortner 2006), who understand social life as a continuously unfolding and iterative process that defies the structure/agency dichotomy. Incorporating such coevolutionary dynamics into EEGs might, however, generate complex interactions that make comparisons between game and real-life behavior even more challenging. Indeed, if the insights of practice theorists are taken seriously, behavior observed in naturalistic settings may be no more or less "real" than behavior observed in experimental settings. EEGs, it seems, embody more theoretical assumptions than game theory alone can supply. Until a shared theory of institutions is articulated and explicitly linked to particular features of EEG design, we may continue arguing past each other in debates about external validity.

4.2. Methodological Issues

Acknowledging both disciplinary tradition and the pragmatic limitations of fieldwork, Chibnik (2005, 202) writes, "anthropologists often explicitly strive to avoid . . . [experimental] manipulation; our goal is to minimize the effects of our presence on what we observe." Surely, anthropologists using EEGs still wish to minimize their impact, but questions about external validity should also raise methodological issues that uniquely or especially concern anthropologists: experimenter demand effects and pedagogical effects.

First, relationships with anthropological-informants-turned-game-participants is an important and overlooked aspect of anthropological engagement with experimental economic methods. Experimenter demand effects arise when participants alter their behavior based on assumptions about the game's purpose and/or the game administrator's expectations. In other words, there might be a game within the EEG where participants try to infer how they "should" behave. Although

experimental economists are aware of this possibility and challenges it poses, we know little about the extent of experimenter demand effects and their implications for EEG interpretation. Some have studied how unintentional cues in the instructions or the game itself may generate experimenter demand effects (Zizzo 2010), but anthropologists may have more sources of influence to consider. Anthropologists typically invest years—even decades—establishing long-term fieldsites, developing connections with communities and rapport with informants. These preexisting relationships—along with the reasonable expectation on the part of informants/participants of a continuing relationship—likely generate unique experimenter demand effects. Furthermore, the informants/participants understanding of the nature of research and its purpose may be quite culturally variable and even idiosyncratic to the nature of the relationship with the ethnographer. How do these relationships influence gameplay, and in turn external validity? Demand effects are not unique to EEGs and are just as likely to create methodological dilemmas for anthropologists using ethnographic methods. However, anthropological uses of EEGs may introduce novel sources of demand effects that warrant further consideration given their methodological assumptions.

Second, mounting evidence of the pedagogical effects of EEGs should give us pause for both ethical and methodological reasons. Participation in EEGs can provide opportunities for reflection and dialogue (Redpath et al. 2018), motivate participants to (re)consider cooperative dilemmas, and produce results that change perceptions about peers (Cárdenas 2009). This learning can spill over into other EEGs (Cárdenas and Carpenter 2005), as well as into real life (Meinzen-Dick et al. 2018; Turiansky 2017). In the examples above, participation in EEGs resulted in learning that had positive impacts outside of the game. These are reassuring findings for applied anthropologists interested in resolving some of the social dilemmas that stymie conservation and development initiatives. But what happens if/when participation in an EEG results in learning that generates negative impacts? After playing a public goods game and a spite game, one participant profusely thanked the game administrators (myself included) for all that they had taught her. I hope she was referring to the public

goods game, rather than the spite game. The pedagogical effects of EEGs may, in contrast, have disturbing methodological implications for academic anthropologists interested in more theoretical questions—especially those who administer EEGs repeatedly with the same population(s). If learning occurs in EEGs, past game participation may influence future EEG results as well as confound analyses of external validity using behaviors observed after gameplay. Regardless of their impact on external validity, both experimenter demand effects and pedagogical effects warrant further consideration and necessitate longitudinal study by anthropologists using EEGs.

On the other hand, anthropologists may be well suited to transform both of these potential methodological liabilities into methodological innovations. More reflexive and/or participatory methods take advantage of close relationships and pedagogical effects to create space for informant-participants to articulate their own understanding of the EEG, reflect on similarities between the game and everyday life, and contextualize game results. For example, after administering framed and unframed PGGs to Siberian fishers and herders, Gerkey (2013) individually interviewed each participant. Contrary to the game's assumptions, many players understood the game in terms of risk rather than trust, and low contributions were perceived as signs of need rather than selfishness. Similarly, Castillo and colleagues (2011) learned more about individual contextual factors influencing gameplay by combining an EEG with multiple methods, even allowing participants to redesign the game through roleplaying exercises. What more might be gleaned by taking seriously informant-participant interpretations of EEG results and perceptions of external validity? Further blending of EEGs with the more qualitative methodologies familiar to anthropologists—such as semi-structured interviews, participatory research designs, and group discussions—is a promising area of new research.

Finally, interpreting the external validity of EEGs raises analytical issues that anthropologists may want to consider. Specifically, what counts as evidence of external validity? While most analyses rely on statistical analyses that provide quantitative evidence, some economists have suggested that

qualitative evidence is a more realistic expectation (Levitt and List 2007b). By qualitative evidence, they mean results consistent with directional or relative predictions that may not meet thresholds for statistical significance. Previous studies have found qualitative evidence of external validity (e.g., Ockenfels and Weimann 1999; Slonim et al. 2013), and this may be appropriate for some studies. For example, imagine a cross-cultural study of generosity using charitable giving as the external indicator. Perhaps correlations and regression analyses do not provide quantitative evidence of parallelism, but rank ordering of countries by generosity and charitable donations results in identical lists. How should such results be interpreted in terms of external validity? The answer likely depends on the particular research goals and questions.

5. CONCLUSIONS

Anthropologists are increasingly turning to EEGs to answer questions about economic decision-making and sociality. Disciplinary differences between anthropology and economics and concerns about external validity, however, can make interpreting EEG results challenging. In this review and reorganization of the literature, I propose generalizability and parallelism as two distinct aspects of external validity. The question of generalizability, on the one hand, highlights the different assumptions that anthropologists and economists have about the universality of preferences. Here the literature is more consistent with anthropological assumptions of variability: EEG results do not reliably generalize across populations, subject pools, settings, or contextual frames. The question of parallelism, on the other hand, reveals different assumptions about the nature of institutions. The empirical record on parallelism is inconsistent, with as many positive as negative results.

Methodological eclecticism has long been a hallmark of anthropology as a discipline, and the anthropological engagement with EEGs should be viewed as a continuation rather than departure from this tradition. In other words, "experimental data can be an important *complement* to observational

and self-report data" (Pisor et al. 2020, 2; emphasis in original), but not a replacement for them. The bare bones of classical EEGs can be given ethnographic flesh to increase context parallelism, but making sense of the results requires more qualitative and descriptive data (Anderies et al. 2011). To this end, scholars of the commons are paying closer attention to how "micro-situational and contextual variables" (Poteete, Janssen, and Ostrom 2010) influence decision-making in EEGs. Anthropologists are therefore uniquely positioned to design better tests of parallelism that incorporate multiple measures of behavior in the future. We should, however, be open to the possibility that variation across cultures and individuals has implications for *both* generalizability *and* parallelism. If the effects of micro-situational and contextual variables are themselves variable, is it possible to design EEGs with high parallelism for all participants in all circumstances? For now, given the variability of analytical strategies and results in the existing literature examining parallelism between game and nongame behavior, we should treat the relationship between EEG results and behavior in daily life as an empirical question for each particular context. Anthropologists have already impacted economics and other disciplines that use EEGs by challenging core assumptions about universal preferences and generalizability. Further engagement with questions of parallelism thus presents an opportunity for anthropologists to develop new insights about external validity and influence how other disciplines in the social sciences study and understand human behavior.

Nicole Naar *Department of Anthropology, University of California, Davis, CA 95616, USA; nanaar@ucdavis.edu; she/her.*

NOTES

Acknowledgments. This research was funded by two grants from the National Science Foundation: NSF-GRFP (Award # 1650042) and NSF-DDRI (Award # 135704). Many thanks to all who participated in the Proseminar on Experimental Economic Games (supported by the UC-Davis

Institute for Social Sciences) for their stimulating ideas and discussion, in particular Travis Lybbert and Peter Richerson. I am also grateful to the members of UC-Davis's EEHBC lab for their valuable comments and feedback on the manuscript, especially Monique Borgerhoff Mulder and Cristina Moya. Finally, this manuscript greatly benefited from the generous, thoughtful, and constructive feedback provided by three wonderful anonymous reviewers.

REFERENCES CITED

Anderies, John M., Marco A. Janssen, François Bousquet, Juan-Camilo Cárdenas, Daniel Castillo, Maria-Claudio Lopez, Robert Tobias, et al. 2011. "The Challenge of Understanding Decisions in Experimental Studies of Common Pool Resource Governance." *Ecological Economics* 70 (9): 1571–79. doi:10.1007/s10683-012-9327-7.

Bauer, Michal, Alessandra Cassar, Julie Chytilová, and Joseph Henrich. 2014. "War's Enduring Effects on the Development of Egalitarian Motivations and In-Group Biases." *Psychological Science* 25 (1): 47–57. doi:10.1177/0956797613493444.

Benz, Matthias, and Stephan Meier. 2008. "Do People Behave in Experiments as in the Field? Evidence from Donations." *Experimental Economics* 11 (3): 268–81. doi:10.1007/s10683-007-9192-y.

Beramendi, Pablo, Raymond Duch, and Akitaka Matsuo. 2016. "Comparing Modes and Samples in Experiments: When Lab Subjects Meet Real People." *SSRN Electronic Journal*. doi:10.2139/ssrn.2840403.

Bouma, Jetske, Erwin Bulte, and Daan van Soest. 2008. "Trust and Cooperation: Social Capital and Community Resource Management." *Journal of Environmental Economics and Management* 56 (2): 155–66. doi:10.1016/j.jeem.2008.03.004.

Bourdieu, Pierre. 1977. *Outline of a Theory of Practice*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511812507.

Brosnan, Sarah F. 2013. "Justice- and Fairness-Related Behaviors in Nonhuman Primates." *Proceedings of the National Academy of Sciences* 110 (Supplement 2): 10416–23. doi:10.1073/pnas.1301194110.

Camerer, Colin F. 2015. "The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List." In *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter, 249–95. Oxford: Oxford University Press. doi:10.2139/ssrn.1977749.

Cameron, Lisa A. 1999. "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia." *Economic Inquiry* 37 (1): 47–59. doi:10.1111/j.1465-7295.1999.tb01415.x.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin.

Cárdenas, Juan-Camilo. 2009. "Experiments in Environment and Development." *Annual Review of Resource Economics* 1:157–82. doi:10.1146/annurev.resource.050708.144056.

Cárdenas, Juan-Camilo, and Jeffrey P. Carpenter. 2005. "Three Themes on Field Experiments and Economic Development." In *Field Experiments in Economics*, edited by Jeffrey P. Carpenter, G. W. Harrison, and John A. List, 71–123. Greenwich: JAI Press. doi:10.1016/S0193-2306(04)10004-5.

Cárdenas, Juan-Camilo, and Jeffrey P. Carpenter. 2008. "Behavioural Development Economics: Lessons from Field Labs in the Developing World." *Journal of Development Studies* 44 (3): 311–38. doi:10.1080/00220380701848327.

Cárdenas, Juan-Camilo, Marco Janssen, and François Bousquet. 2013. "Dynamics of Rules and Resources: Three New Field Experiments on Water, Forests and Fisheries." In *Handbook on Experimental Economics and the Environment*, edited by John A. List and Michael K. Price, 319–45. Northampton: Edward Elgar Publishers. doi:10.4337/9781781009079.00020.

Carpenter, Jeffrey P., Stephen V. Burks, and Eric A. Verhoogen. 2005. "Comparing Students to Workers: The Effects of Social Framing on Behavior in Distribution Games." *Research in Experimental Economics* 10:261–90. doi:10.1016/s0193-2306(04)10007-0.

Carpenter, Jeffrey, and Erika Seki. 2011. "Do Social Preferences Increase Productivity? Field Experimental Evidence from Fishermen in Toyama Bay." *Economic Inquiry* 49 (2): 612–30. doi:10.1111/j.1465-7295.2009.00268.x.

Cassar, Alessandra, Feven Wordofa, and Y. Jane Zhang. 2016. "Competing for the Benefit of Offspring Eliminates the Gender Gap in Competitiveness." *Proceedings of the National Academy of Sciences* 113 (19): 5201–5. doi:10.1073/pnas.1520235113.

Castillo, Daniel, François Bousquet, Marco A. Janssen, Kobchai Worrapimphong, and Juan Camilo Cardenas. 2011. "Context Matters to Explain Field Experiments: Results from Colombian and Thai Fishing Villages." *Ecological Economics* 70 (9): 1609–20. doi:10.1016/j.ecolecon.2011.05.011.

Chibnik, Michael. 2005. "Experimental Economics in Anthropology: A Critical Assessment." *American Ethnologist* 32 (2): 198–209. doi:10.1525/ae.2005.32.2.198.

Cookson, R. 2000. "Framing Effects in Public Goods Experiments." *Experimental Economics* 79:55–79. doi:10.1007/bf01669207.

Cronk, Lee. 2007. "The Influence of Cultural Framing on Play in the Trust Game: A Maasai Example." *Evolution and Human Behavior* 28 (5): 352–58. doi:10.1016/j.evolhumbehav.2007.05.006.

Cronk, Lee, and Helen Wasielewski. 2008. "An Unfamiliar Social Norm Rapidly Produces Framing Effects in an Economic Game." *Journal of Evolutionary Psychology* 6 (4): 283–308. doi:10.1556/JEP.6.2008.4.3.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2): 448–74. doi:10.1257/jel.47.2.448.

Druckman, James N., and Cindy D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Data Base.'" In *Cambridge Handbook of Experimental Political Science*, edited by James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia, 41–57. New York: Cambridge University Press. doi:0.1017/CBO9780511921452.004.

Ellingsen, Tore, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. 2012. "Social Framing Effects: Preferences or Beliefs?" *Games and Economic Behavior* 76 (1): 117–30. doi:10.1016/j.geb.2012.05.007.

Englmaier, Florian, and Georg Gebhardt. 2016. "Social Dilemmas in the Laboratory and in the Field." *Journal of Economic Behavior and Organization* 128:85–96. doi:10.1016/j.jebo.2016.03.006.

Ensminger, Jean. 2002. "Experimental Economics: A Powerful New Method for Theory Testing in Anthropology." In *Theory in Economic Anthropology*, edited by Jean Ensminger, 59–78. Walnut Creek: Alta.

Fehr, Ernst, and John A. List. 2004. "The Hidden Costs and Returns of Incentives: Trust and Trustworthiness among CEOs." *Journal of the European Economic Association* 2 (5): 743–71. doi:10.2139/ssrn.364480.

Fiedler, Marina, and Ernan Haruvy. 2009. "The Lab versus the Virtual Lab and Virtual Field: An Experimental Investigation of Trust Games with Communication." *Journal of Economic Behavior & Organization* 72:716–24. doi:10.1016/j.jebo.2009.07.013.

Franzen, Axel, and Sonja Pointner. 2013. "The External Validity of Giving in the Dictator Game: A Field Experiment Using the Misdirected Letter Technique." *Experimental Economics* 16 (2): 155–69. doi:10.1007/s10683-012-9337-5.

Fréchette, Guillaume R. 2015. "Laboratory Experiments: Professionals versus Students." In *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter, 360–90. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780195328325.003.0019.

Frigau, Luca, Tiziana Medda, and Vittorio Pelligra. 2019. "From the Field to the Lab: An Experiment on the Representativeness of Standard Laboratory Subjects." *Journal of Behavioral and Experimental Economics* 78:160–9. doi:10.1016/j.socec.2018.06.003.

Gächter, Simon. 2010. "(Dis)Advantages of Student Subjects: What Is Your Research Question?" Working Paper Series Des Rates Für Sozial-Und Wirtschaftsdaten, Berlin. doi:10.1017/s0140525x10000099.

Galizzi, Matteo M., and Daniel Navarro-Martinez. 2019. "On the External Validity of Social Preference Games: A Systematic Lab-Field Study." *Management Science* 65 (3): 955–1453. doi:10.1287/mnsc.2017.2908.

Gelcich, Stefan, Ricardo Guzman, Carlos Rodriguez-Sickert, Juan Carlos Castilla, and Juan-Camilo Cárdenas. 2013. "Exploring External Validity of Common Pool Resource Experiments: Insights from Artisanal Benthic Fisheries in Chile." *Ecology and Society* 18 (3): 2. doi:10.5751/es-05598-180302.

Gerkey, Drew. 2013. "Cooperation in Context: Public Goods Games and Post-Soviet Collectives in Kamchatka, Russia." *Current Anthropology* 54 (2): 144–76. doi:10.1086/669856.

Gerlach, Philipp. 2017. "The Games Economists Play: Why Economics Students Behave More Selfishly than Other Students." *PLoS ONE* 12 (9): e0183814. doi:10.1371/journal.pone.0183814.

Gerlach, Philipp, and Bastian Jaeger. 2016. "Another Frame, Another Game? Explaining Framing Effects in Economic Games." In *Proceedings of Norms, Actions, Games*, edited by A. Hopfensitz and E. Lori. Toulouse: Institute for Advanced Studies. doi:10.17605/OSF.IO/AB5YP.

Gervais, Matthew M. 2017. "RICH Economic Games for Networked Relationships and Communities: Development and Preliminary Validation in Yasawa, Fiji." *Field Methods* 29 (2): 113–29. doi:10.1177/1525822X16643709.

Giddens, Anthony. 1986. *The Constitution of Society: Outline of the Theory of Structuration*. Berkeley: University of California Press.

Gintis, Herbert. 2006. "The Foundations of Behavior: The Beliefs, Preferences, and Constraints Model." *Biological Theory* 1 (2): 123–27. doi:10.1162/biot.2006.1.2.123.

Goeschl, Timo, Sara Elisa Kettner, Johannes Lohse, and Christiane Schwieren. 2015. "What Do We Learn from Public Good Games about Voluntary Climate Action? Evidence from an Artefactual Field Experiment." *Discussion Paper Series*, No. 595. University of Heidelberg, Department of Economics. doi:10.2139/ssrn.2620229.

Gurven, Michael, Arianna Zanolini, and Eric Schniter. 2008. "Culture Sometimes Matters: Intra-Cultural Variation in Pro-Social Behavior among Tsimane Amerindians." *Journal of Economic Behavior and Organization* 67 (3–4): 587–607. doi:10.1016/j.jebo.2007.09.005.

Hagen, Edward H., and Peter Hammerstein. 2006. "Game Theory and Human Evolution: A Critique of Some Recent Interpretations of Experimental Games." *Theoretical Population Biology* 69 (3): 339–48. doi:10.1016/j.tpb.2005.09.005.

Handberg, Øyvind Nystad, and Arild Angelsen. 2015. "Experimental Tests of Tropical Forest Conservation Measures." *Journal of Economic Behavior & Organization* 118:346–59. doi:10.1016/j.jebo.2015.03.007.

Henrich, Joseph. 2000. "Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining among the Machiguenga of the Peruvian Amazon." *The American Economic Review* 90 (4): 973–79. doi:10.1257/aer.90.4.973.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, et al. 2005. "'Economic Man' in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies." *Behavioral and Brain Sciences* 28 (6): 795–815; discussion 815–55. doi:10.1017/S0140525X05000142.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2–3): 61–83; discussion 83–135. doi:10.1017/S0140525X0999152X.

Henrich, Joseph, Richard Mcelreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, et al. 2006. "Costly Punishment across Human Societies." *Science* 312 (5781): 1767–71. doi:10.1126/science.1127333.

Hodgson, Geoffrey M. 2006. "What Are Institutions?" *Journal of Economic Issues* XL (1): 1–25. doi:10.1080/00213624.2006.11506928.

Hooghe, Marc, Dietlind Stolle, Valérie-Anne Mahéo, and Sara Vissers. 2010. "Why Can't a Student Be More Like an Average Person? Sampling and Attrition Effects in Social Science Field and

Laboratory Experiments." *The Annals of the American Academy of Political and Social Science* 628 (March): 85–96. doi:10.1177/0002716209351516.

Jackson, Cecile. 2012. "Internal and External Validity in Experimental Games: A Social Reality Check." *European Journal of Development Research* 24 (1): 71–88. doi:10.1057/ejdr.2011.47.

Javaid, Aneeque, Micaela M. Kulesz, Achim Schlüter, Alexandra Ghosh, and Narriman S. Jiddawi. 2016. "Time Preferences and Natural Resource Extraction Behavior: An Experimental Study from Artisanal Fisheries in Zanzibar." *PLoS ONE* 11 (12): 1–14. doi:10.1371/journal.pone.0168898.

Lesorogol, Carolyn K. 2007. "Bringing Norms In: The Role of Context in Experimental Dictator Games." *Current Anthropology* 48 (6): 920–26. doi:10.1086/523017.

Lesorogol, Carolyn K. 2017. "Fairness in Cultural Context." In *Interdisciplinary Perspectives on Fairness, Equity, and Justice*, edited by Meng Li and David P. Tracer, 129–42. Cham: Springer International Publishing. doi:10.1007/978-3-319-58993-0.

Levitt, Steven D., and John A. List. 2007a. "Viewpoint: On the Generalizability of Lab Behaviour to the Field." *Canadian Journal of Economics* 40 (2): 347–70. doi:10.1111/j.1365-2966.2007.00412.x.

Levitt, Steven D., and John A. List. 2007b. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21 (2): 153–74. doi:10.1257/jep.21.2.153.

Levitt, Steven D, and John A. List. 2008. "Homo Economicus Evolves." *Science* 319 (5865): 909–10. doi:10.1126/science.1153640.

Loewenstein, George. 1999. "Experimental Economics from the Vantage Point of Behavioural Economics." *The Economic Journal* 109 (453): 25–34. doi:10.1111/1468-0297.00400.

Luo, Jiayi, and Rongjun Yu. 2015. "Follow the Heart or the Head? The Interactive Influence Model of Emotion and Cognition." *Frontiers in Psychology* 6 (May). doi:10.3389/fpsyg.2015.00573.

Macfarlan, Shane J., and Robert J. Quinlan. 2008. "Kinship, Family, and Gender Effects in the Ultimatum Game." *Human Nature* 19 (3): 294–309. doi:10.1007/s12110-008-9045-1.

Meinzen-Dick, Ruth, Marco A. Janssen, Sandeep Kandikuppa, Rahul Chaturvedi, Kaushalendra Rao, and Sophie Theis. 2018. "Playing Games to Save Water: Collective Action Games for Groundwater Management in Andhra Pradesh, India." *World Development* 107:40–53. doi:10.1016/j.worlddev.2018.02.006.

Ockenfels, Axel, and Joachim Weimann. 1999. "Types and Patterns: An Experimental East-West-German Comparison of Cooperation and Solidarity." *Journal of Public Economics* 71:275–87. doi:10.1016/s0047-2727(98)00072-3.

Oosterbeek, Hessel, Randolph Sloof, and Gijs van de Kuilen. 2004. "Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis." *Experimental Economics* 7:171–88. doi:10.2139/ssrn.286428.

Ortner, Sherry B. 2006. *Anthropology and Social Theory: Culture, Power, and the Acting Subject*. Durham: Duke University Press. doi:10.1215/9780822388456.

Paciotti, Brian, and Craig Hadley. 2003. "The Ultimatum Game in Southwestern Tanzania: Ethnic Variation and Institutional Scope." *Current Anthropology* 44 (3): 427–32. doi:10.1086/374903.

Pisor, Anne C., Matthew M. Gervais, Benjamin G. Purzycki, and Cody T. Ross. 2020. "Preferences and Constraints: The Value of Economic Games for Studying Human Behaviour." *Royal Society Open Science* 7:192090. doi:10.1098/rsos.192090.

Pisor, Anne C., and Michael Gurven. 2018. "When to Diversify, and with Whom? Choosing Partners among Out-Group Strangers in Lowland Bolivia." *Evolution and Human Behavior* 39 (1): 30–39. doi:10.1016/j.evolhumbehav.2017.09.003.

Poteete, Amy R., Marco Janssen, and Elinor Ostrom. 2010. *Working Together: Collective Action, the Commons, and Multiple Methods in Practice*. Princeton: Princeton University Press. doi:10.1515/9781400835157.

Prediger, Sebastian, Björn Vollan, and Markus Frolich. 2011. "The Impact of Culture and Ecology on Cooperation in a Common-Pool Resource Experiment." *Ecological Economics* 70 (9): 1599–608. doi:10.1016/j.ecolecon.2010.08.017.

Purzycki, Benjamin Grant, Coren Apicella, Quentin D. Atkinson, Emma Cohen, Rita Anne McNamara, Aiyana K. Willard, Dimitris Xygalatas, et al. 2016. "Moralistic Gods, Supernatural Punishment and the Expansion of Human Sociality." *Nature* 530 (7590): 327–30. doi:10.1038/nature16980.

Redpath, Steve M., Aidan Keane, Henrik Andrén, Zachary Baynham-Herd, Nils Bunnefeld, A. Bradley Duthie, Jens Frank, et al. 2018. "Games as Tools to Address Conservation Conflicts." *Trends in Ecology and Evolution* 33 (6): 415–26. doi:10.1016/j.tree.2018.03.005.

Roe, Brian E., and David R. Just. 2009. "Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments, and Field Data." *American Journal of Agricultural Economics* 91 (5): 1266–71. doi:10.1111/j.1467-8276.2009.01295.x.

Rommel, Jens. 2015. "What Can Economic Experiments Tell Us about Institutional Change in Social-Ecological Systems?" *Environmental Science and Policy* 53:96–104. doi:10.1016/j.envsci.2014.05.006.

Rucas, Stacey L., Michael Gurven, Hillard Kaplan, and Jeffrey Winking Jr. 2010. "The Social Strategy Game: Resource Competition within Female Social Networks among Small-scale Forager-Horticulturalists." *Human Nature* 21 (1): 1–18. doi:10.1007/s12110-010-9079-z.

Schram, Arthur. 2005. "Artificiality: The Tension between Internal and External Validity in Economic Experiments." *Journal of Economic Methodology* 12 (2): 225–37. doi:10.1080/13501780500086081.

Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51 (3): 515–30. doi:10.1037//0022-3514.51.3.515.

Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy and Public Affairs* 6 (4): 317–44. doi:10.2307/2264946.

Shapley, Harlow. 1964. *Of Stars and Men: Human Response to an Expanding Universe*. Westport: Greenwood Press.

Slonim, Robert, Carmen Wang, Ellen Garbarino, and Danielle Merrett. 2013. "Opting-In: Participation Bias in Economic Experiments." *Journal of Economic Behavior and Organization* 90:43–70. doi:10.1016/j.jebo.2013.03.013.

Smith, Eric Alden. 2005. "Making It Real: Interpreting Economic Experiments." *Behavioral and Brain Sciences* 28 (6): 832–33. doi:10.1017/s0140525x0540014x.

Smith, Vernon L. 1982. "Microeconomic Systems as an Experimental Science." *The American Economic Review* 72 (5): 923–55. doi:10.1017/cbo9780511528354.018.

Stoop, Jan, Charles N. Noussair, and Daan van Soest. 2012. "From the Lab to the Field: Cooperation among Fishermen." *Journal of Political Economy* 120 (6): 1027–56. doi:10.1086/669253.

Sullivan, Roger J., and Henry F. Lyle III. 2005. "Economic Models Are Not Evolutionary Models." *Behavioral and Brain Sciences* 28 (6): 836. doi:10.1017/s0140525x05430149.

Thaler, Richard H. 2000. "From Homo Economicus to Homo Sapiens." *Journal of Economic Perspectives* 14 (1): 133–41. doi:10.4324/9781315106045-6.

Torres-Guevara, Luz Elba, and Achim Schlüter. 2016. "External Validity of Artefactual Field Experiments: A Study on Cooperation, Impatience and Sustainability in an Artisanal Fishery in Colombia." *Ecological Economics* 128:187–201. doi:10.1016/j.ecolecon.2016.04.022.

Turiansky, Abbie. 2017. "Collective Action in Games as in Life: Experimental Evidence from Canal Cleaning in Haiti." Working Paper 57. Oakland: Mathematica Policy Research. https://www.mathematica-mpr.com/our-publications-and-findings/publications/collective-action-in-games-as-in-life-experimental-evidence-from-canal-cleaning-in-haiti.

Winking, Jeffrey, and Nicholas Mizer. 2013. "Natural-Field Dictator Game Shows No Altruistic Giving." *Evolution and Human Behavior* 34 (4): 288–93. doi:10.1016/j.evolhumbehav.2013.04.002.

Zizzo, Daniel John. 2010. "Experimenter Demand Effects in Economic Experiments." *Experimental Economics* 13:75–98. doi:10.1007/s10683-009-92.

FIGURE CAPTIONS

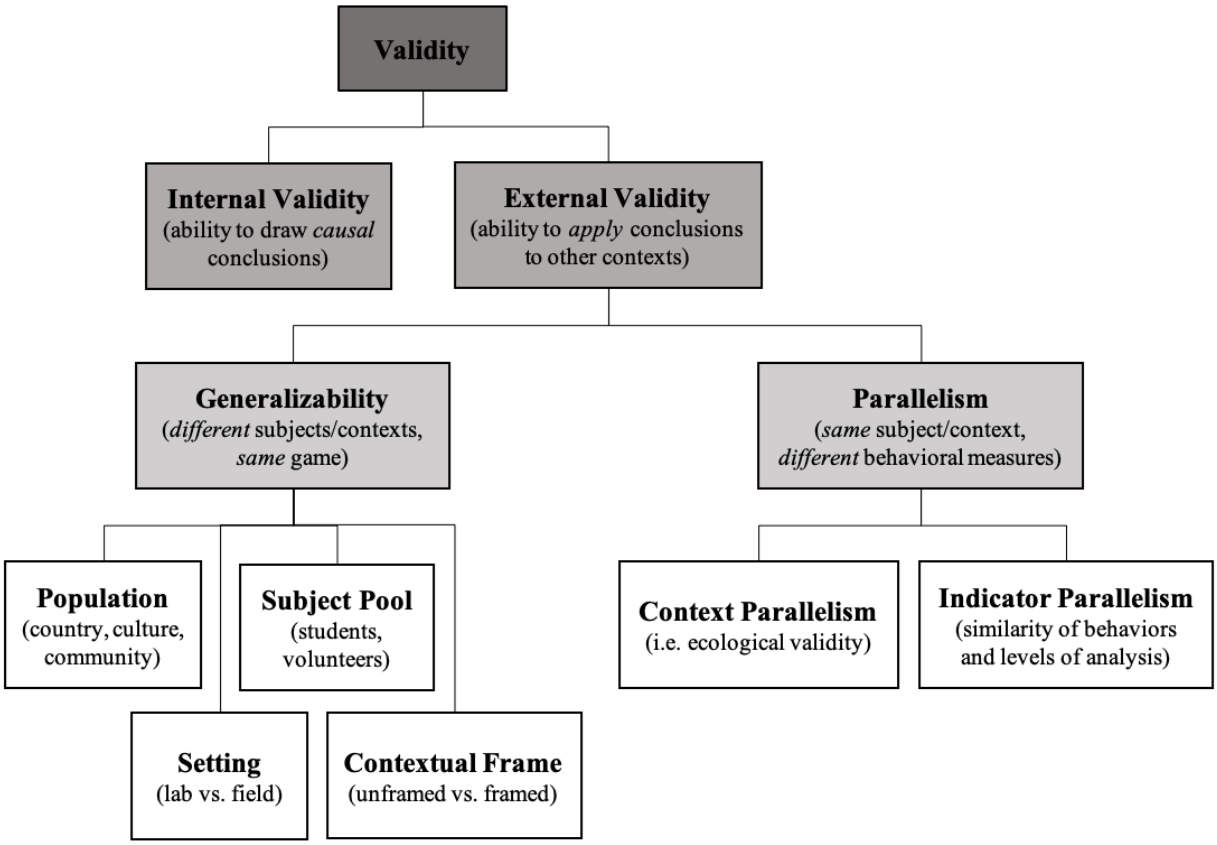


Figure 1. *Different kinds of validity that inform the interpretation of experimental economic games.*

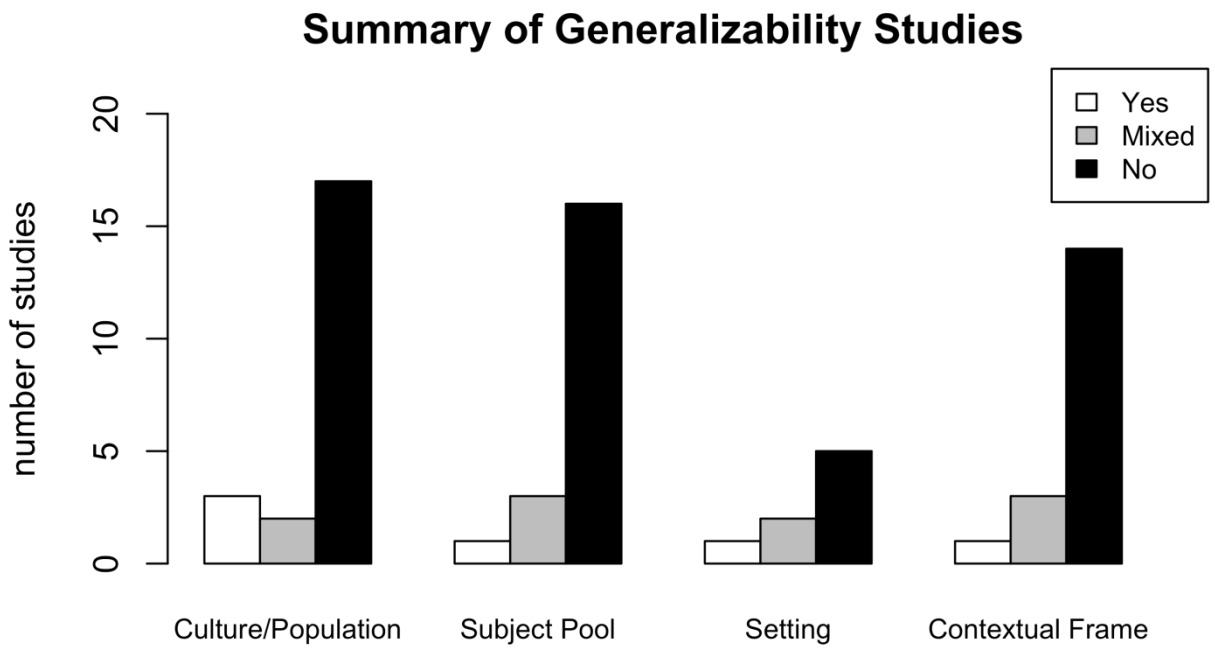


Figure 2. *A summary of the previous studies of generalizability reviewed in Tables S1–S4 of the Supplementary Materials. Each study was classified—based on the conclusions reached by the study's author(s)—as having positive (Yes), mixed (Mixed), or negative (No) evidence of generalizability.*

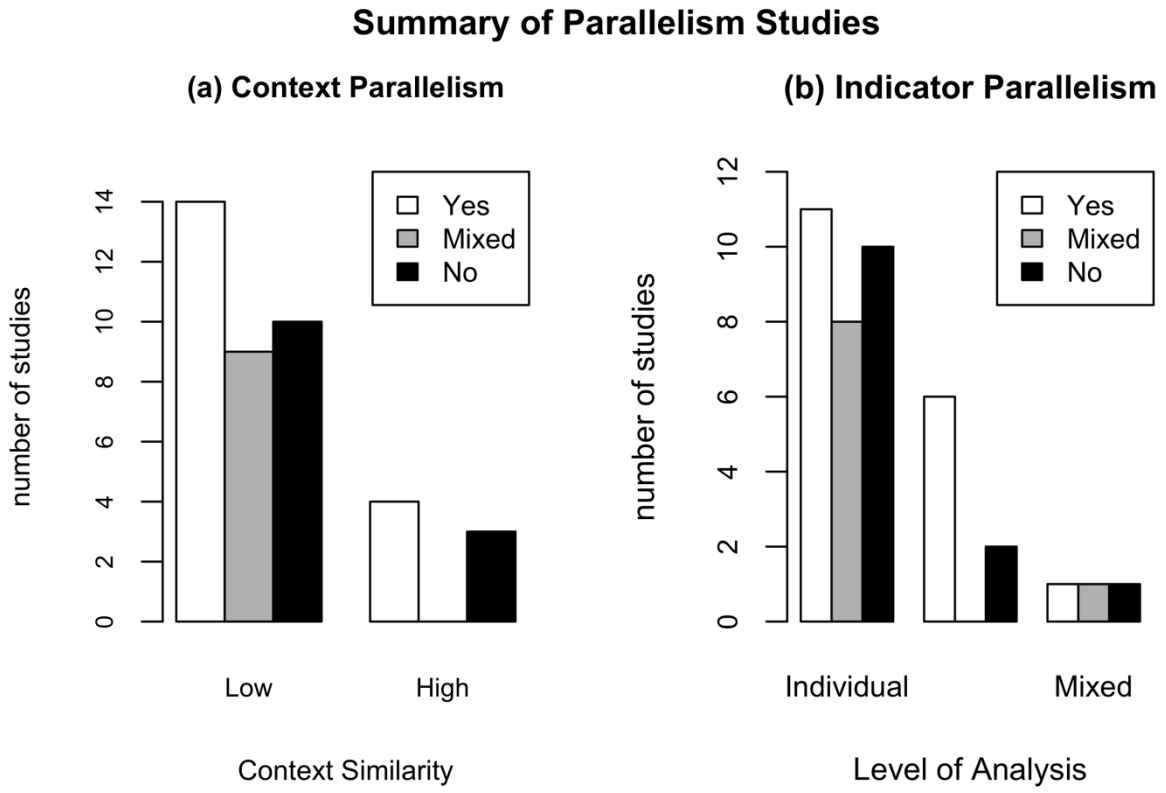


Figure 3. *A summary of the previous studies of parallelism reviewed in Tables S5 and S6 of the Supplementary Materials. Each study was classified—based on the conclusions reached by the study's author(s)—as having positive (Yes), mixed (Mixed), or negative (No) evidence of parallelism.*