PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

On the reliability of individual economic rationality measurements

Felix J. Nitsch[a,b,c,1], Luca M. Lüpken[a], Nils Lüschow[a], and Tobias Kalenscher[a]

Edited by Tobias Hauser, University College London, London, United Kingdom; received February 5, 2022; accepted June 23, 2022 by Editorial Board Member Michael S. Gazzaniga

A contemporary research agenda in behavioral economics and neuroeconomics aims to identify individual differences and (neuro)psychological correlates of rationality. This research has been widely received in important interdisciplinary and field outlets. However, the psychometric reliability of such measurements of rationality has been presumed without enough methodological scrutiny. Drawing from multiple original and published datasets (in total over 1,600 participants), we unequivocally show that contemporary measurements of rationality have moderate to poor reliability according to common standards. Further analyses of the variance components, as well as a allowing participants to revise previous choices, suggest that this is driven by low between-subject variance rather than high measurement error. As has been argued previously for other behavioral measurements, this poses a challenge to the predominant correlational research designs and the search for sociodemographic or neural predictors. While our results draw a sobering picture of the prospects of contemporary measurements of rationality, they are not necessarily surprising from a theoretical perspective, which we outline in our discussion.

rationality | reliability | econometrics | psychometrics | measurement

Significance

Identifying potential determinants of rationality—interpreted as a characteristic of decision makers—is of great relevance from an applied science perspective: both policy makers and industry have a pronounced interest in understanding which individuals make rational decisions, be it to design effective policies, enhance equity, or fine-tune talent selection processes. However, especially for research at the frontier of foundation to application, we must ensure that our measurements are precise and reliable. Here, we show that established empirical measurements of rationality are not reliable enough, implicating the urgent need for advances in measurement of rationality.

A common definition of economic rationality states that decision makers should consistently choose the subjectively best option according to their preferences as their budget allows. It can be shown that any collection of choices of such decision makers can be reconciled with a definite structure of wants, cost efficiency, and transitivity (1). Specifically, the generalized axiom of revealed preference (GARP) requires that if a decision maker accepts costs to choose one choice object over another (strict direct revealed preference), they may, *ceteris paribus,* in fact never choose the latter over the former choice object (no direct revealed preference) as long as it is not associated with higher costs (2–4). Rational choice theory is theoretically parsimonious and elegant and delivers tractable analytical results; therefore, it is widely used in economic applications. However, since its inception, this standard model has also received severe criticism on descriptive (e.g., 5), predictive (e.g., 6), and normative (e.g., 7) grounds.

Despite this criticism, a contemporary research agenda in psychology, behavioral economics, and neuroeconomics aims to identify individual differences and (neuro)psychological correlates of rationality [(8, 9) for an overview]. Here, indices of revealed preference consistency specifically are used as an ad hoc measurement tool for the supposedly latent concept of rationality that is often interpreted as a psychological construct (see *Discussion*). Importantly, this interpretation of revealed preference consistency as a characteristic of decision makers goes beyond the original intent of the founders of revealed preference theory, which was to provide a test of whether a specific set of choices allows for the construction of a preference ordering (e.g., 10). Empirical research on this topic has been widely received in important interdisciplinary and field outlets such as *Science* or the *American Economic Review* (e.g., 11, 12). However, the validity of such measurements of rationality has been, perhaps due to the strong economic-theoretical foundation, presumed without enough scrutiny.

In this article, we identify and discuss a core issue of the aforementioned research program: contemporary measurements of individual rationality have moderate to poor reliability according to common standards. As has been argued previously for other behavioral measurements (13), this poses a challenge for the predominant correlational research designs and the identification of individual differences.

Importantly, the empirical analyses reported in this paper draw from multiple original and published datasets that vary in the deployed choice domain (social choice, food choice, choice under risk, or choice under ambiguity), choice complexity (two or three goods), study context (laboratory or online), incentivization (incentive compatible or

Author affiliations: [a]Comparative Psychology, Institute of Experimental Psychology, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany; [b]Marketing Area, INSEAD, Fontainebleau, France; and [c]Control-Interoception-Attention Team, Paris Brain Institute, INSERM U 1127, CNRS UMR 7225, Sorbonne University, Paris, France

Author contributions: methodology, formal analysis, data curation, writing - original draft, visualization, project administration by F.J.N.; conceptualization by F.J.N., L.M.L., and T.K.; software by F.J.N., L.M.L., and N.L.; supervision by F.J.N. and T.K.; investigation by F.J.N. and L.M.L.; writing - review and editing by F.J.N., L.M.L., N.L., and T.K.; and funding acquisition by T.K.

The authors declare no competing interest.

This article is a PNAS Direct Submission. T.H. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: felixjan.nitsch@insead.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2202070119/-/DCSupplemental.

Published July 26, 2022.

hypothetical), study population, sample size, task structure, measurement length, and time gap between measurements (*SI Appendix*, Table S1). Hence, the large amount of data and methodological diversity allow us to draw conclusions with reasonable generality for contemporary research practice.* To support the robustness of our results, we collected data and replicated the low reliability of rationality measurements in eight datasets with, in total, over 1,600 participants, including a preregistered replication.

In comparison to previous related work on the reliability of choice consistency (e.g., 14, 15), the revealed preference methodology for the measurement of rationality is the standard in research practice (8) and is not only conceptually but also mathematically linked to utility theory (1). Another feature that distinguishes revealed preference consistency from choice consistency more generally is the consideration of consumer theory, specifically price effects on demand (16). Thus, our investigation specifically focuses on the measurements of economic rationality via revealed preferences.

A clear implication of our research is that the reliability of such rationality measurements cannot be assumed until shown otherwise. While few (perhaps none) of the relevant studies in the field report reliability coefficients, our results suggest that reliability is modest even for more conservative study designs. More broadly, however, we ask how valid rational choice theory is as a measurement model for differential-psychological applications.

Results

Analysis Approach. Following the standard approach appropriated from neoclassic economic theory, rationality was quantified via the two most prominent indices of rationality, namely, Afriat's critical cost efficiency index [CCEI (17)] and the Houtman-Maks index [HMI (18, 19)]. Broadly, the CCEI utilizes the fact that irrational choice behavior is not cost efficient. It denotes the minimal hypothetical waste of wealth that a decision maker accepts given their irrationality.† The HMI, on the other hand, does not consider the fraction of wealth wasted but instead determines the size of the largest subset of choices consistent with GARP. Hence, the HMI is possibly more robust to single outliers (e.g., mistake choices) but also more sensitive to multiple but practically negligible violations of rationality (see *Methods* for further details). To quantify the reliability of rationality measurements, we calculated the intraclass correlation coefficients (ICCs), which we report and interpret following the standards put forth by Koo and Li [(21); preregistered for study 2], where "[v]alues less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability."

Study 1: Initial Finding. In our first online study, we included 53 adult, English-speaking participants recruited via the online platform Prolific. Participants solved two measurements (test and retest) of a modified dictator game with 20 trials each (22). Briefly, in each trial, participants were granted a variable monetary endowment. Importantly, they could freely share a fraction of this endowment with their best friend at variable exchange rates (i.e., sometimes their friend could receive more or less money than given up by the participant). This design allowed us to determine the rationality of their revealed preferences in sharing.

Importantly, we manipulated the way the dictator game was presented: we used two common design versions (Fig. 1 A and B) and a novel version (Fig. 1C). Each participant solved all 2 (test vs. retest) × 3 (design) = 6 measurement-task version combinations.

The number and configuration of trials was confirmed to be sufficient to detect violations of rationality via a task-based power analysis (see *Rationality and reliability*). Between the two measurements, participants solved an unrelated filler task on reading comprehension (Fig. 1D and *SI Appendix*). At the end of the experiment, participants answered several questions regarding their decision strategies and experiences solving the tasks.

Rationality and reliability. To determine the statistical power of our GARP test, we bootstrapped 1,000 virtual participants from our dataset as a model of random choice (23). Results showed that Bronars power = 91.8% bootstrapped participants did not comply with GARP (see *Methods*, axiom), indicating that the task can accurately detect random behavior. Overall, the rationality of our actual participants (quantified by either CCEI or HMI) was relatively high for both measurements (*SI Appendix*, Table S1) and significantly higher than a bootstrapped random benchmark of equal sample size (all $P < 0.001$; *SI Appendix*, Fig. 13).

Intermethod reliability between task versions. The intermethod reliability (between task versions) for the CCEI was ICC (2, 1) = 0.071 (−0.108 < ICC < 0.297) for the first measurement and ICC (2, 1) = 0.356 (95% CI: 0.176 < ICC < 0.539) for the second measurement (Fig. 2, *Top*). Similarly, the intermethod reliability (across task versions) for the HMI was ICC (2, 1) = 0.094 (−0.089 < ICC < 0.320) for the first measurement and ICC (2, 1) = 0.309 (95% CI: 0.129 < ICC < 0.497) for the second measurement (Fig. 2, *Bottom*). Hence, the intermethod reliability of the CCEI and the HMI was poor for both measurements according to common standards.

Test-retest reliability per task version. The test-retest reliability (within task versions) for the CCEI was ICC (2, 1) = 0.626 (95% CI: 0.404 < ICC < 0.779) for the diagram task, ICC (2, 1) = 0.439 (95% CI: 0.180 < ICC < 0.641) for the bundles task, and ICC (2, 1) = 0.277 (95% CI: −0.021 < ICC < 0.531) for the slider task. Overall, only the diagram task showed moderate test-retest reliability for the CCEI, while the two other tasks performed poorly (Fig. 3, *Left*). The test-retest reliability (within task versions) for the HMI was ICC (2, 1) = 0.345 (95% CI: 0.054 < ICC < 0.583) for the diagram task, ICC (2, 1) = 0.550 (95% CI: 0.317 < ICC < 0.720) for the bundles task, and ICC (2, 1) = 0.310 (95% CI: 0.014 < ICC < 0.556) for the slider task. Overall, only the bundles task showed moderate test-retest reliability for the HMI, while the two other tasks performed poorly (Fig. 3, *Right*).

Decision strategies. In order to gain a better understanding of the decision-making process and to further validate our conclusions, we conducted an inductive, qualitative content analysis (24) using the free-text responses about the decision strategies of our participants (*SI Appendix*, Fig. 1A). Our results indicated that most participants either tried to fairly share the payout (22 participants, 41.50%) or maximize the total payout (20 participants, 37.70%). Few participants decided with an egotistical bias (6 participants, 11.30%) or prosocial bias (2 participants, 3.77%). For 3 participants (5.66%), we could not determine a clear strategy from their response.

*We, however, acknowledge that this diversity also introduces heterogeneity, which limits the means of quantitative data aggregation.

†For our main analysis, we quantified this hypothetical waste in terms of percentages of the monetary expenditure. To test the robustness of our results, we used an alternative specification in terms of the absolute waste of expenditure (20). The results of this robustness check are aligned with our main analysis (*SI Appendix*).

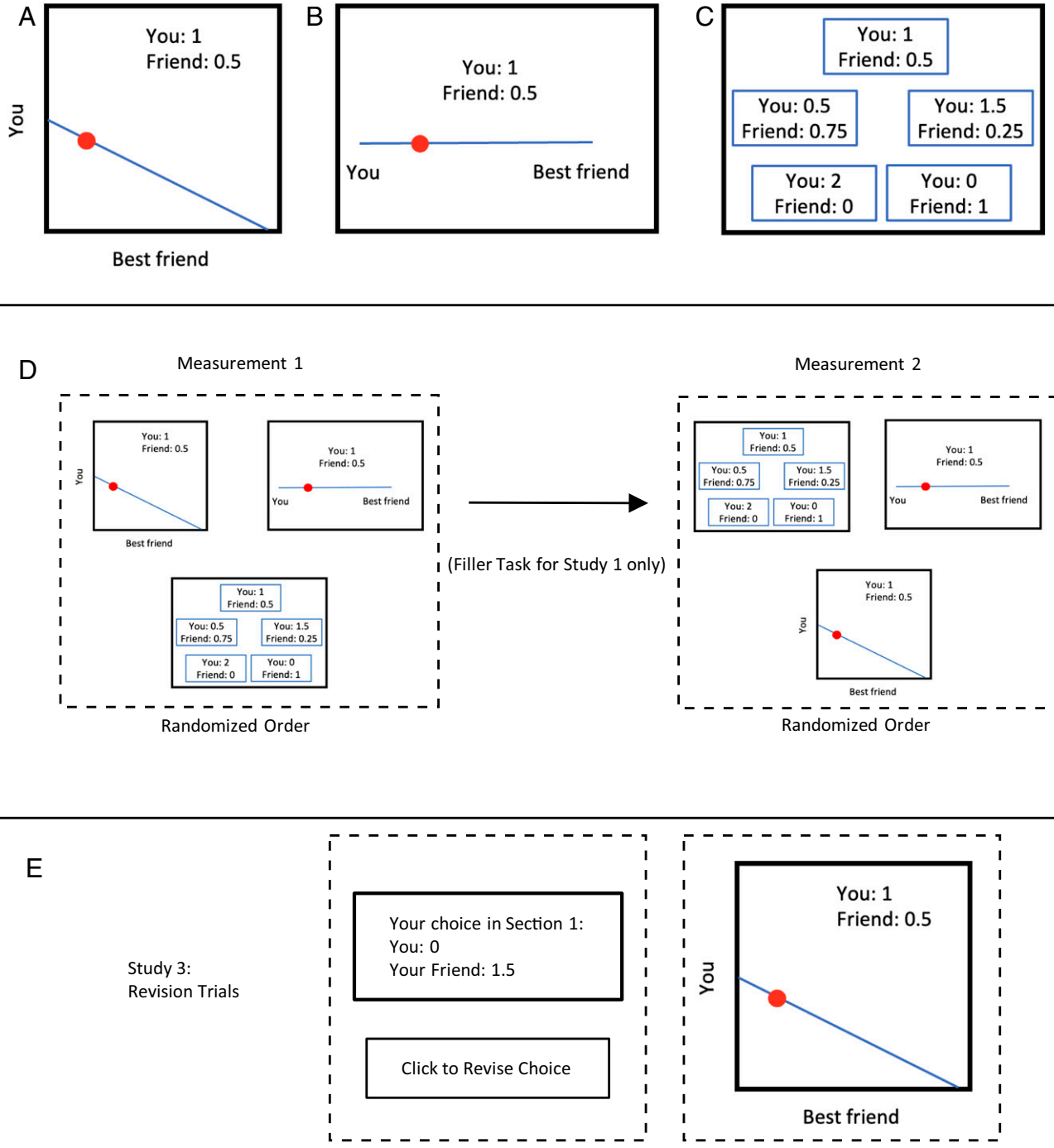

Fig. 1. Behavioral task versions used to measure rationality. Participants played a modified dictator game. In each trial, they could share a fraction of an endowment with their best friend at a variable exchange rate (i.e., sometimes their friend could receive more or less money than given up by the participant). We manipulated the way the decision problem was presented in three different task versions: (*A*) diagram, (*B*) slider, and (*C*) bundles. (*D*) Experimental structure of studies 1 and 2. All task version blocks were presented in randomized order for two measurements each. Intermethod reliability was calculated within measurements (across task versions). Test-retest reliability was calculated between measurements (per task version). (*E*) Sample trial of the choice revisions in study 3. Participants first were displayed their choice from the previous section and then had the opportunity to either remake or revise that decision.

Study 2: Preregistered Replication. A potential limitation of study 1 was the relatively small number of participants and trials,[‡] which could have led to unstable and biased reliability estimates (25). To address this concern, we conducted study 2, which was preregistered on the Open Science Framework (OSF, https://osf.io/wfd4z). Here, we tried to replicate the results of study 1 in a larger sample and with a higher number of trials. We included 148 adult, English-speaking Prolific participants who did not participate in study 1. Participants underwent the same procedure as in study 1 except for two differences. First, we increased the number of trials from 20 to 40. The number and configuration of trials were confirmed to be sufficient to detect violations of rationality via a task-based power analysis (see *Rationality and reliability*). Second, we omitted the filler task to compensate for the higher number of trials and thus longer experiment duration. At the end of the experiment, participants answered several questions regarding their decision strategies and experiences solving the tasks.

[‡]The number of trials for study 1 was confirmed to be sufficient via a Bronars power analysis and well within the range used in the literature.

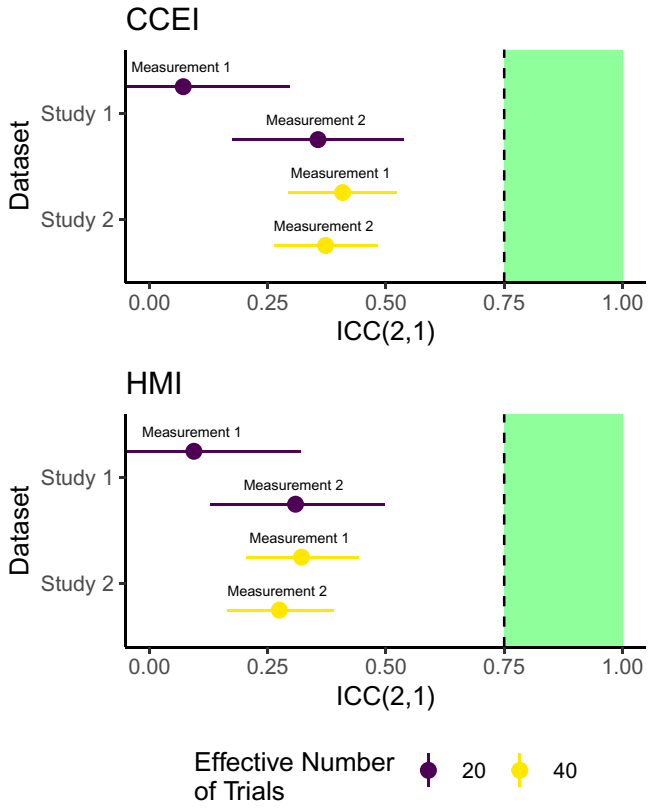

Fig. 2. Intermethod reliability of individual rationality measurements. Depicted are ICC estimates and 95% CI of the intermethod reliability of CCEI (*Top*) and HMI (*Bottom*). The dashed vertical line and subsequent green area indicate the range of acceptable, that is, good reliability according to common standards. The effective number of trials is the number of trials per measurement (test-retest reliability).

Rationality and reliability. As in study 1, to determine the statistical power of our GARP test, we bootstrapped 1,000 virtual participants from our dataset. Results showed that Bronars power > 99.9% bootstrapped participants did not pass GARP, indicating that the task can accurately detect random behavior. Again, the rationality of our participants (quantified by either CCEI or HMI) was relatively high for both measurements (*SI Appendix*, Table S1) and significantly higher than a bootstrapped random benchmark (all $P < 0.001$; *SI Appendix*, Fig. 14).

Intermethod reliability between task versions. The intermethod reliability (between task versions) for the CCEI was ICC (2, 1) = 0.408 (95% CI: 0.293 < ICC < 0.522) for the first measurement and ICC (2, 1) = 0.372 (95% CI: 0.263 < ICC < 0.482) for the second measurement. Hence, as in study 1, the intermethod reliability of the CCEI was poor for both measurements (Fig. 2, *Top*). The intermethod reliability (between task versions) for the HMI was ICC (2, 1) = 0.321 (95% CI: 0.204 < ICC < 0.442) for the first measurement and ICC (2, 1) = 0.275 (95% CI: 0.164 < ICC < 0.390) for the second measurement. Again, the intermethod reliability of the HMI was poor for both measurements (Fig. 2, *Bottom*).

Test-retest reliability per task version. The test-retest reliability (within task versions) for the CCEI was ICC (2, 1) = 0.515 (95% CI: 0.372 < ICC < 0.635) for the diagram task, ICC (2, 1) = 0.497 (95% CI: 0.354 < ICC < 0.617) for the bundles task, and ICC (2, 1) = 0.434 (95% CI: 0.283 < ICC < 0.564) for the slider task. Overall, test-retest reliability for the CCEI was moderate for the diagram and bundles tasks and poor for the slider task; therefore, it was not sufficient according to our preregistered criterion (at least good reliability; Fig. 3, *Left*). The test-retest reliability (within task versions) for the HMI was ICC (2, 1) = 0.505 (95% CI: 0.360 < ICC < 0.626) for the diagram task, ICC (2, 1) = 0.640 (95% CI: 0.525 < ICC < 0.732) for the bundles task, and ICC (2, 1) = 0.343 (95% CI: 0.182 < ICC < 0.487) for the slider task. Overall, test-retest reliability for the HMI was moderate for the diagram and bundles tasks and poor for the slider task following standards; therefore, it was not sufficient according to our preregistered criterion (at least good reliability; Fig. 3, *Right*).

Decision strategies. Free-text responses about participants' decision strategy were categorized via the same categories inducted in study 1 (*SI Appendix*, Fig. 1B). Again, our results showed that most participants either tried to fairly share the payout (63 participants, 42.6%) or maximize the total payout (43 participants, 29.1%). Few participants decided with an egotistical bias (25 participants, 16.9%) or prosocial bias (3 participants, 2.0%). For 14 participants (9.5%), we could not determine a clear strategy from their response. This number was expectedly higher than for study 1 as no new response categories were inducted.

Reliability of Rationality in Published Research. Next, to assess the reliability of contemporary rationality measurements more generally, we reanalyzed five published datasets for test-retest or split-half reliability: Choi et al. [(26), henceforth C07, and (11), henceforth C14], Kurtz-David et al. [(27), henceforth K19], the control group of Nitsch et al. [(28, 29), henceforth N21], and Ahn et al. [(30), henceforth A14]. *SI Appendix*, Table S1 shows key details about the datasets. The N21 dataset contained three measurements over about 3 h in total, which were jointly entered for the reliability estimation. Since the C07, C14, K19, and A14 datasets did not contain multiple measurements, we used the split-half method to estimate reliability.

Split-half/test-retest reliability. Results for all datasets are summarized in Fig. 3. The split-half/test-retest reliability for the CCEI was ICC (2, 1) = 0.256 (95% CI: 0.056 < ICC < 0.436) for the C07 dataset, ICC (2, 1) = 0.503 (95% CI: 0.402 < ICC < 0.591) for the C14 dataset, ICC (2, 1) = 0.183 (95% CI: −0.166 < ICC < 0.491) for the K19 dataset, ICC (2, 1) = 0.483 (95% CI: 0.340 < ICC < 0.619) for the N21 dataset, and ICC (2, 1) = 0.408 (95% CI: 0.268 < ICC < 0.532) for the A14 dataset (Fig. 3, *Left*, rows 6–10). Hence, split-half/test-retest reliability for the CCEI was poor in four of five datasets and moderate in one of five datasets. The split-half/test-retest reliability for the HMI was ICC (2, 1) = 0.442 (95% CI: 0.263 < ICC < 0.592) for the C07 dataset, ICC (2, 1) = 0.158 (95% CI: 0.032 < ICC < 0.279) for the C14 dataset, ICC (2, 1) = 0.685 (95% CI: 0.451 < ICC < 0.831) for the K19 dataset, and ICC (2, 1) = 0.497 (95% CI: 0.355 < ICC < 0.630) for the N21 dataset (Fig. 3, *Right*, rows 6–9).[§] Hence, split-half/test-retest reliability for the HMI was poor in three of four datasets and moderate in one of four datasets.

Study 3: Long-Term Reliability and the Role of Mistake Choices. Study 3 was a follow-up measurement within a subset of the participants of study 2 ∼5 mo later, which had three main goals. The first goal was to replicate our findings in participants who were already familiar with the task. This was important since we observed learning effects in studies 1 and 2, where participants became on average more rational in the second measurement [study 1: b = 0.023, SE = 0.017, $t(240.932) = 1.387$, $P = 0.167$; study 2: b = 0.037, SE = 0.02, $t(658.251) = 3.183$, $P = 0.002$]. The second goal was to explore how test-retest

[§]We failed to reproduce the HMI for the A14 dataset with our algorithm, presumably due to the presence of three different kinds of goods.

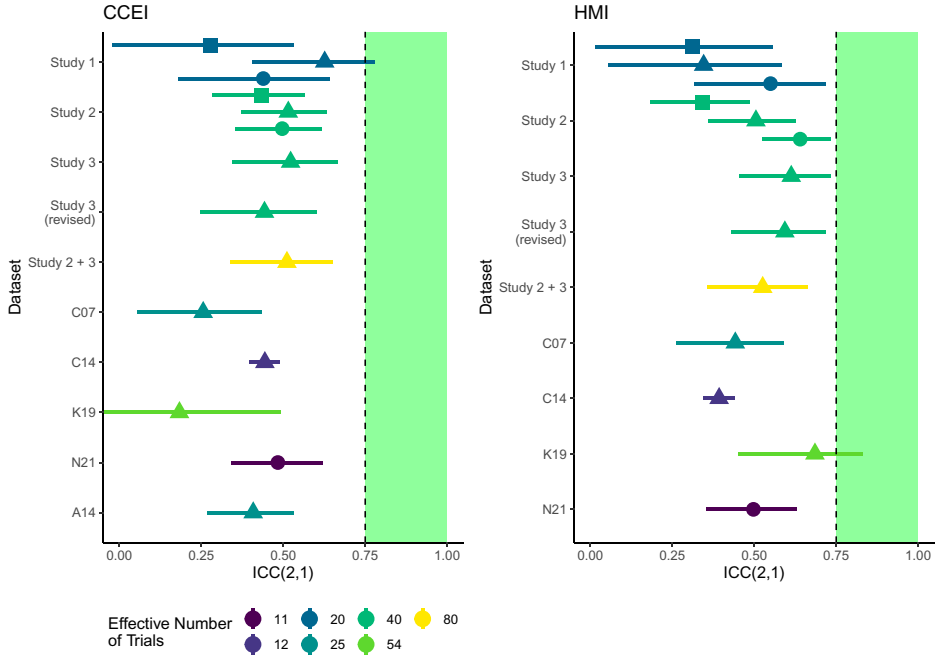

Fig. 3. Test-retest and split-half reliability of individual rationality measurements. Depicted are the ICC estimates and 95% CI of the test-retest/split-half reliability of CCEI (*Left*) and HMI (*Right*) across all eight datasets. Symbols indicate which task version was used: triangles indicate the diagram task, circles indicate the bundles task, and rectangles indicate the slider task. The dashed vertical line and subsequent green area indicate the range of acceptable, that is, good reliability according to common standards. The effective number of trials is the number of trials per measurement (test-retest reliability) or split (split-half reliability). Studies 2 and 3 were conducted ~5 mo apart in the same sample.

reliability was affected when the two tests were further apart (months rather than minutes). For this, we leveraged the fact that we recruited a subset of the sample of study 2 and the same diagram choice task. The third goal was to test whether low reliability was driven by the noisiness of participants' decisions (i.e., mistake choices). We were able to rerecruit 97 of the original 148 participants.[¶] Participants underwent the same procedure as in study 2 except for three differences. First, we omitted the slider and bundles tasks and only used the diagram task (for a total of 2 × 40 trials). Second, we omitted the free-text questions regarding participants' decision strategy (to limit both the length of the study and fatigue in participants; see *Discussion*). Third, following Breig and Feldman (31), we allowed participants to revise a random subset of their initial choices after the completion of the first two task blocks and undo potential mistake choices.

Rationality and reliability. Bronars power was equivalent to study 2, that is, >99.9% bootstrapped participants did not pass GARP, indicating that the task can accurately detect random behavior. Again, the rationality of our participants (quantified by either CCEI or HMI) was relatively high for both measurements (*SI Appendix*, Table S1) and significantly higher than a bootstrapped random benchmark (all $P < 0.001$; *SI Appendix*, Figs. 15 and 16).

Goal 1: Reliability of participants who are familiar with the task. The test-retest reliability for the CCEI was ICC (2, 1) = 0.522 (95% CI: 0.343 < ICC < 0.665), which was comparable to study 2 and can be considered moderate. The test-retest reliability for the HMI was ICC (2, 1) = 0.613 (95% CI: 0.455 < ICC < 0.733), which was slightly higher than for study 2 but still only moderate.

Goal 2: Long-term test-retest reliability. We collapsed the 2 × 40 trials of studies 2 and 3 each for an effective number of 80 trials per study. The 5-month test-retest reliability (across studies 2 and 3) for the CCEI was ICC (2, 1) = 0.511 (95% CI: 0.338 < ICC < 0.651), which can be considered moderate. The 5-month test-retest reliability (across studies 2 and 3) for the HMI was ICC (2, 1) = 0.526 (95% CI: 0.355 < ICC < 0.662), which can be considered moderate. Hence, overall the 5-month test-retest reliability of both indices was comparable to short-term reliability.

Goal 3: The role of mistake choices. Following Breig and Feldman (31), we allowed participants to revise a random subset of 10 choices per block of their initial choices after the completion of the first two task blocks and undo potential mistake choices. Surprisingly, this led to an increase neither of rationality (measurement 1: delta mean CCEI = −0.025, delta mean HMI = −0.262; measurement 2: delta mean CCEI = −0.013, delta mean HMI = −0.058) nor of test-retest reliability. The test-retest reliability for the CCEI was ICC (2, 1) = 0.443 (95% CI: 0.248 < ICC < 0.603), which can be considered poor. The test-retest reliability for the HMI was ICC (2, 1) = 0.593 (95% CI: 0.431 < ICC < 0.719), which can be considered moderate.

Explorative Analysis of Variance Components. Next, we tried to answer the question of what gives rise to the low reliability of the revealed preference indices. As explained above, the ICC represents the fraction of the total variance which is attributable to true differences (i.e., not attributable to error). Hence, the ICC could be small due to high measurement error (large denominator), small true differences (small numerator), or both. To identify the degree of measurement error, we calculated within-subject coefficient of variance (WSCV) using the repeated measures per the same instrument. The WSCV determines the degree of closeness of the repeated observations made on the same subject (32) (Fig. 4); the lower the WSCV, the lower the measurement error. Results indicated a drop of the WSCV for measurements with at least 20 trials for both CCEI and HMI (i.e., all datasets but C14 and N21). For such

[¶]Study 3 was an unplanned follow-up motivated by a reviewer comment. Hence, we had to rerecruit participants via Prolific who had not been informed about this follow-up in advance.

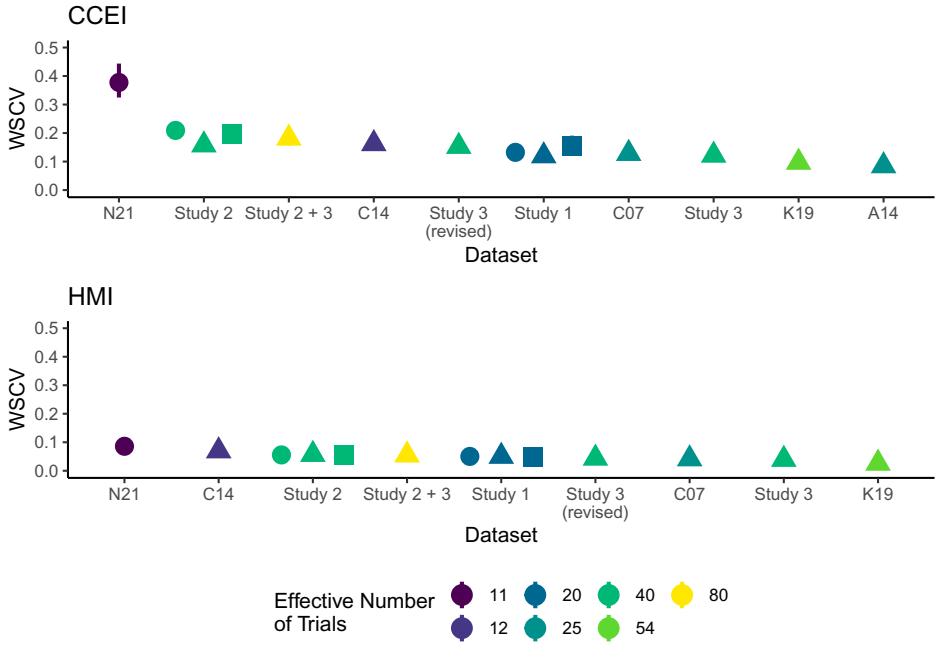

Fig. 4. WSCV. Depicted are the WSCV estimates and 95% CI of the test-retest/split-half reproducibility of CCEI (*Top*) and HMI (*Bottom*) across all eight datasets. Symbols indicate which task version was used: triangles indicate the diagram task, circles indicate the bundles task, and rectangles indicate the slider task. Studies 2 and 3 were conducted ~5 mo apart in the same sample. The effective number of trials is the number of trials per measurement (test-retest reliability) or split (split-half reliability).

measurements with at least 20 trials, the WSCV was relatively small (median = 15% for the CCEI and median = 5% for the HMI).

Discussion

In the present paper, we investigated the reliability of behavioral measurements of rationality as a characteristic of individual decision makers. Across multiple original and published datasets, we found that the reliability of the two most prominent rationality indices (and variations thereof) is moderate to poor. This result held independent of the choice domain (social choice, food choice, choice under risk, or choice under ambiguity), choice complexity (two or three goods), study context (laboratory or online), incentivization (incentive compatible or hypothetical), study population, sample size, task structure, measurement length, and time gap between measurements. Hence, given data from multiple datasets with sufficient methodological diversity, our conclusions not only apply to a specific configuration of rationality measurements but speak with reasonable generality for contemporary research practice. More broadly, our results align with recent work on the reliability of measurements of risk preferences, sensitivity to losses, and self-regulation (33–35).[#]

Reliability indicates how much of the total variance in the variable of interest is attributable to true difference and not caused by measurement error. Hence, one potential explanation of moderate to poor reliability could be the presence of high measurement error (large denominator in the fraction) in the revealed preference methodology (e.g., 38). Our data offer two arguments against this explanation. First, allowing participants to revise a subset of choices (i.e., fixing potential mistakes, a source of measurement error; see study 3) did not increase reliability. Second, an analysis of the variance components in the data tentatively suggested that within-subject variance, as a proxy for measurement error, was sufficiently low for measurements with at least 20 trials.

Another explanation for moderate to poor reliability is a lack of true differences between participants: it could be possible that it is difficult to distinguish between participants because they do differ enough with respect to economic rationality. In line with this explanation, most participants across all datasets descriptively behaved with high consistency, and taking individual measurements of CCEI and HMI yielded approximately two times worse predictive accuracy for another measurement within the same individual than simply assuming the population mean (*SI Appendix*). In conjunction with the absence of high measurement error, this tentatively suggests that the low reliability of contemporary measurements of individual rationality (that is, the inability to distinguish between individuals) was indeed driven by a lack of interindividual differences in rationality.[||]

As has been argued previously for other behavioral measurements, the lack of reliability poses a challenge to the contemporary search for sociodemographic or psychological correlates of economic rationality. Pragmatically speaking, our results show that a simple increase of trials or using a different task interface is not sufficient to fix this problem (unless the sample size is increased substantially); rather, individual differences must be increased. Possible avenues to explore here are, for instance, to ask participants to make decisions under stress or time pressure, increasing the difficulty of the decisions or using a manipulation (i.e., a between-groups design).

A more general point we want to raise here is the validity of economic rationality as a psychological construct, which is a prerequisite to valid measurements (39). Strictly speaking, economic rationality describes whether or to what extent a set of choices can be described by a utility function. In the recent literature,

[#]Interestingly, temporal discounting seems to be reliable (36, 37).

[||]It is important to point out that few individual differences in rationality pose a desirable result for economic theory: specifically, it means that most individuals' decisions can be closely approximated by utility theory.

however, economic rationality of a finite set of choices of an individual has been compared with decision quality (11, 29, 40), policy responsiveness (41), or variability in the neural computation of value (27) based on face validity and correlational evidence,** all of which are arguably related but not identical constructs. Given this lack of a clear definition of the psychological construct to be measured and lack of evidence of validity for widely used measures, advances in psychological theory and measurement appear necessary. A recent, particularly promising approach for the applications outlined here are generative models,†† which can serve to formalize psychological constructs and increase the reliability of behavioral measurements (43, 44). The theoretical basis to inspire the development of such models could be provided by cognitive science: past studies have shown that cognitive skills (e.g., executive control, working memory, and intelligence) pose a common factor of many aspects of decision making, including choice consistency (45–50).

Limitations. A limitation to our assessment might be that the upper limit of effective trials considered for our reliability assessments was 80, which is below that of a few particularly high-powered studies (e.g., [27] used up to 108 trials). Optimistically, it could be possible to improve the reliability of rationality measurements by increasing the number of trials in behavioral experiments to further reduce measurement error. However, a high number of trials (i.e., at least more than 80 trials) comes at the cost of practical feasibility in many studies and the risk of increasing fatigue due to prolonged measurement durations, which in itself might bias the measurement (ranging from relative changes in preference to qualitative changes of decision strategy). To give a benchmark, in study 2, the number of participants indicating that they were "very much fatigued" tripled from after 20 to after 40 trials (after measurement 1: $n = 10$; after measurement 2: $n = 29$; *SI Appendix*, Fig. 2).

Another potential limitation is that we could not replicate the finding of Breig and Feldman (31) that allowing participants to revise their choices leads to an increase of revealed preference consistency in the choice set. We acknowledge that, as the authors also demonstrate in their paper, the effectiveness of such an intervention depends on the specific configuration of the choice interface. Hence, due to the ineffectiveness of the intervention, we cannot rule out that a more effective intervention could increase the reliability of rationality measurements.

Lastly, we acknowledge that while the qualitative results for each dataset are similar, there is some variability in the quantitative reliability estimates, which could be driven by the heterogeneity of the included datasets.

Conclusions. We demonstrate that the reliability of individual rationality measurements cannot be assumed until shown otherwise. While few (perhaps none) of the relevant studies in the field report reliability coefficients, our results suggest that reliability is modest even for more conservative study designs. From the theoretical perspective outlined above, however, we might ask more broadly how useful a measurement model rational choice theory (or choice structure representations thereof) is for differential-psychological applications.

**A notable exception is Cohen et al. (42), who provide mechanistic and causal evidence for the link between neuronal constraints and economic rationality in nematodes.

††Generative models are models that formally specify "how behavior is generated within people and how generative processes vary across people" (43, p. 2).

Methods

All participants recruited for this research project gave their informed written consent before participation. The study protocol of the original studies 1 to 3 was approved by the ethical council of the medical faculty of Heinrich-Heine-University Düsseldorf (study 2020-910). Studies 1 to 3 were conducted in alignment with the Declaration of Helsinki. For ethical information regarding the literature data, see the corresponding references. Study 1 was conducted as part of the doctoral thesis of F.J.N. [51].

Study 1. Study 1 served as the initial investigation into the reliability of measures of revealed preference reliability and was embedded in a larger study on the malleability of rational choice.

Participants. For study 1, 101 adult, English-speaking participants completed our study. For the study, 48 participants were randomly assigned to an experimental manipulation group, which entailed a reading-based priming manipulation. As this manipulation was irrelevant to the presented research question, we only considered the control group (which only read a neutral text; *SI Appendix*) for the present analyses. No other participants were excluded, resulting in a final sample size of $n = 53$ participants. *SI Appendix*, Table S2 gives an overview of the demographics.

Procedure and design. Participants were recruited via the online platform Prolific (https://www.prolific.co), receiving compensation of 4.30 pounds. Prolific is a widely used online research subject pool that has been accredited for more transparency and research suitability than comparable platforms (52, 53). The online experiment was programmed in jsPsych (54) and hosted on Pavlovia. Before the start of the experiment, all participants were fully debriefed about the content and aim of the research project and provided informed consent via a checkbox. After providing consent, we asked for their demographic information. Next, participants underwent the first measurement of all three experimental tasks in randomized order. For the first measurement, each task entailed a detailed description and five practice trials. After completion of the first measurement, participants solved a filler task that consisted of reading three informational texts about unrelated topics and answering three quiz questions on the content of these texts (*SI Appendix*). Then participants underwent the second measurement of all three experimental tasks, again in randomized order. At the end of the experiment, participants answered several questions regarding their decision strategies and experiences solving the tasks. Then they were redirected back to Prolific to receive their compensation.

Our experimental design was completely within subject. Participants solved all three decision tasks for two measurements (3 × 2 within-subject design).

Experimental tasks. All decision tasks were based on a modified dictator game (22), consisting of $I = 20$ decisions per measurement. Participants had to hypothetically allocate a budget m_i between them and their best friend, resulting in a final monetary split of $x_i = (x_i^{Self}, x_i^{Friend})$. Importantly, the monetary endowment m_i and the "prices" of keeping and giving money $p_i = (p_i^{Self}, p_i^{Friend})$ varied per decision. Hence, $x_i^{Self} = \frac{share_i^{Self} m_i}{p_i^{Self}}$ and $x_i^{Friend} = \frac{share_i^{Friend} m_i}{p_i^{Friend}}$, with the share indicating the relative fraction of the budget (0–1) allocated to each account. Budgets and prices were randomly sampled per trial: $m_i \in [2, 3, 4, 5, 6, 7, 8, 9, 10]$ and $p_i^{Self}, p_i^{Friend} \in [1, 2, 3]$. For our analysis, we normalized prices and budgets so that $\sum p_i = 1$ and $m_i = x_i^{Self} p_i^{Self} + x_i^{Friend} p_i^{Friend}$.

For each task and measurement, we further included two attention check trials where participants were instructed to allocate the full budget to either themselves or their best friend. Those trials were not included in the analysis. If participants failed an attention check for a given measurement of a task, we excluded that measurement of the task from our analysis specifically (8% of measurements).

Contemporarily, experimental investigations of revealed preference rationality interchangeably utilize different ways to present the decision problems (task versions). One line of research uses a task introduced by Choi et al. (26). In their elaborate and widely used paradigm (here, diagram task), participants must allocate a budget between two dimensions (e.g., two investment accounts, oneself and a coplayer) using a cartesian coordinate display. The task is mostly applied in the investigation of choices under risk (11, 27, 55, 56) but also intertemporal choices (12, 57). It has the appeal that it transparently depicts all economic parameters (budget, prices, budget line, etc.) and even allows for a visual

identification of inconsistent choices. A potential drawback is that the task can be hard to understand for people without experience in the interpretation of diagrams and the theoretically large number of potential choice options.

Another line of research uses a more simplistic choice-bundles task that was first prominently used by Harbaugh and colleagues (58) and by many others since (40, 41, 59). In the choice-bundles task (here, bundles task), the budget line is divided into (equidistant) discrete points, which are subsequently presented as a discrete set of choice options to the participant. Conveniently, in this task, participants can ignore the underlying economic parameters and must only choose the most liked choice bundle in the set. This significantly reduces the cognitive demand of the task and is desirable for indivisible goods (e.g., food items), specific participant groups (e.g., children), and research questions (e.g., decisions under stress). A drawback of the task is that discrete choice options can only approximate optimal choices from a continuous budget line, which might introduce small inconsistencies by itself.

A compromise suggested by Garagnani (60), which so far has not been widely applied, however, would be to present participants with a slider (here, slider task) that allows for the continuous allocation of the budget while concealing economic parameters to a degree that allows for an intuitive approach to solving the task. Concretely, participants move a slider that controls the fraction of money allocated to each dimension (in our study, self and best friend). The effective payouts are shown in an infobox.

In study 1, we used all task variants in order to be able to draw inferences independent of the specific task design and evaluate intermethod reliability (Fig. 1).

Diagram task. For each decision, participants had to choose a point on a diagonal line in a coordinate system (Fig. 1A). The points on the diagonal line represented the possible money allocations between themselves and their best friend that they might choose. In each coordinate system, the vertical axis corresponded to the money chosen for themselves (you) and the horizontal axis corresponded to the money chosen for their best friend (friend). While they were making their decision, they could see which amount of money they had chosen for themselves and for their best friend in the upper right corner of the coordinate system. The flatter the lines, the more money their best friend could receive as a maximum compared to them. The steeper the lines, the more money they could receive as a maximum compared to their best friend.

Bundles task. For each decision, participants had a choice of five different money allocations and were instructed to simply choose the allocation that they thought was best (Fig. 1C).

Slider task. For each decision, participants had to choose a point on a horizontal line by moving a slider, which represented the possible allocations of money amounts between themselves and their best friend. While making their decision, they could see which amount of money they had chosen for themselves and their best friend in two boxes above the slider. The labeling of the endpoints and spatial presentation were randomized from round to round (Fig. 1B).

Task questionnaires. At the end of the experiment, participants were asked to answer how they reached their decisions ("How did you reach your decisions?") and what they considered particularly important in their decisions ("What was particularly important to you in your decisions?" in study 1 only) in open-text format. Further, they were asked multiple questions regarding their experiences with the specific task formats that will be reported elsewhere.

Study 2. Study 2 was preregistered on OSF (https://osf.io/wfd4z/). Here, we tried to replicate the results of study 1 in a larger sample and with a higher number of trials to address the issue of portability, as this theoretically could increase the reliability of the task.

Participants. We recorded complete data of 148 adult, English-speaking participants, none of which were excluded. Our sample size rationale was based on the maximum feasible sample size given our monetary budget and the second-largest and largest non–panel-based sample of all datasets considered [*SI Appendix*, Table S1, following Lakens (61)]. *SI Appendix*, Table S2 gives an overview of the demographics.

Procedure and design. Our procedure and design were similar to the control condition of study 1 except that we increased the number of trials from 20 to 40 and removed the filler task. Again, participants were recruited via Prolific, receiving compensation of 4.30 pounds. Participants underwent two measurements of all three experimental tasks, each measurement in randomized order. For the first measurement, each task entailed a detailed description and five practice trials. As study 2 took on average longer than study 1 to complete, we asked participants to indicate their fatigue once after the first measurement and once after the second measurement. The fatigue measurement was used to evaluate the extended experiment length after the recruitment of 10 initial participants, as indicated in our preregistration. Of these 10 participants, most participants were not or only a little fatigued throughout the experiment; therefore, we deemed the experimental length acceptable. At the end of the experiment, participants again answered several questions regarding their decision strategies and experiences solving the tasks. Our experimental design was completely within subject. Participants solved all three decision tasks for two measurements (3×2 within-subject design). Again, if participants failed an attention check for a given measurement of a task, we excluded that measurement of the task from our analysis specifically (10% of measurements).

Study 3. Study 3 was a follow-up measurement within a subset of the participants of study 2 ~5 months later, which served to replicate our findings in participants who were already sufficiently familiar with the task, explore how test-retest reliability was affected when the two tests were further apart (months rather than minutes), and test whether low reliability was driven by the noisiness of participants' decisions (i.e., mistake choices).

Participants. We were able to rerecruit and include 97 of the original 148 participants. *SI Appendix*, Table S2 gives an overview of the demographics.

Procedure and design. Participants underwent the same procedure as in study 2 except for three differences. First, we omitted the slider and bundles tasks and only used the diagram task (for a total of 2×40 trials). Second, we omitted the free-text questions regarding participants' decision strategy (to limit both the length of the study and fatigue in participants; see *Discussion*). Third, following Breig and Feldman (31), we allowed participants to revise a random subset of 10 choices each of their initial choices after the completion of the first two task blocks.

Again, participants were recruited via Prolific, receiving compensation of 3 pounds. Participants underwent two measurements of the diagram task. For the first measurement, we again displayed a detailed description and five practice trials. After the completion of the two first measurements, participants were informed that they now had the opportunity to revise a selection 10 of their decisions from the first (second) measurement (Fig. 1E). For each potential revision, the previous choice was displayed to the participants as a reminder. Then, participants could proceed to redo their choice in their own pace. Importantly, the starting point of the slider on the budget line was again randomized (transparently to the participants) to facilitate an active decision.

Again, if participants failed an attention check for a given measurement of a task, we excluded that measurement of the task from our analysis specifically (12% of measurements).

Analysis.

Revealed preference analysis. Let N be the number of different commodity types in a commodity bundle. Let X be the nonnegative, N-dimensional space of commodity bundles. Let P be the strictly positive, N-dimensional space of prices of commodities. Let M be the nonnegative, one-dimensional space of budgets. Let $I = i, j, \ldots, n$ denote observations of choice. Let x_i be the chosen commodity bundle of an observation i. Each bundle x_i is a N-dimensional vector of the shape $x_i = (x_i^1, x_i^2 \ldots, x_i^n)$, with each scalar component x_i^n representing the quantity of commodity type n within bundle x_i. Let p_i be the given prices of commodities of an observation i. Each prices p are a N-dimensional vector of the shape $p_i = (p_i^1, p_i^2, \ldots p_i^n)$, with each scalar component p_i^n representing the price of commodity type n per unit size. Then the scalar product $x_i \cdot p_i$ represents the total price of a commodity bundle x_i at some prices p_i. Let m_i be the given budget of an observation i. We assume that decision makers spend all their budget so that $x_i \cdot p_i = m_i$.

Definition 1 (Direct Revealed Preference): A bundle x_i is directly revealed preferred to another bundle x_j if and only if $x_j \cdot p_i \leq m_i$. Then we denote $x_i R_D\, x_j$.

Definition 2 (Revealed Preference): A bundle x_i is revealed preferred to another bundle x_k if there exists a transitive preference relation $x_i R_D\, x_j\, R_D\, x_k$ between both bundles. We denote $x_i R x_k$.

Definition 3 (Strict Direct Revealed Preference): A bundle x_i is strictly directly revealed preferred to another bundle x_j if and only if $x_j \cdot p_i < m_i$. Then we denote $x_i P_D\, x_j$.

Axiom (GARP). $x_i R x_j \Leftrightarrow \neg\, x_j P_D\, x_i$.

In this framework, the CCEI presents a relaxation of definition 3 so that only $x_i P_D\, x_j \leftarrow x_j \cdot p_i < CCEI \cdot m_i$ is required.[‡‡] The CCEI is then the highest possible value so that the axiom holds for all observations. The HMI, on the other hand, denotes the largest number of observations $HMI \leq I$ for which the axiom holds.

Statistical analysis. To quantify the reliability of measurements of rationality, we calculated the ICCs of type consistency between the variables of interest to account for potential learning effects [62]. Reliability indicates how much of the total variance in the variable of interest is not caused by measurement error. A reliability of 1 indicates a perfect measurement: all variance in the measurement is caused by true differences in the variable of interest. As reliability approaches 0, the measurement is becoming less precise and a higher fraction of variance is due to measurement error [63]. Specifically, we calculated the ICC of consistency measures among tasks for each measurement to assess intermethod reliability. Further, we calculated the ICC of consistency measures within tasks across measurements to assess test-retest reliability. For qualitative interpretation of reliability, we followed the standards by Koo and Li (21): an ICC estimate lower than 0.50 is considered poor, one between 0.50 and 0.75 is considered moderate, one between 0.75 and 0.90 is considered good, and lastly one greater than 0.90 is considered excellent.

Qualitative content analysis. In order to gain a better understanding of the decision-making process and to further validate our data, we conducted an inductive, qualitative content analysis (24) using the free-text responses about the decision strategies of our participants. For study 1, we concatenated the answers of our participants to both questions. Using an ad hoc inductive approach on the material from study 1, five exhaustive and mutually exclusive categories were generated. Subsequently, three raters (of which one was blind to the objective of the study) coded all answers independently for studies 1 and 2 using these categories without further revisions. On their first try, our raters reached good interrater reliability for study 1 [ICC (2, 1) = 0.831; 95% CI: 0.750 < ICC < 0.892] and moderate interrater reliability study 2 [ICC (2, 1) = 0.740; 95% CI: 0.676 < ICC < 0.796]. The final categorization was then decided via majority rule where possible or labeled "unclear" otherwise.

Data Availability. Processed data and anonymized raw data and analysis scripts have been deposited in OSF (https://osf.io/KD4HW/). Previously published data were used for this work (11,26–30,64). All other study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. Portions of this work were developed from the doctoral thesis of F.J.N. We thank Paul Kramer for support with the qualitative content analysis, as well as Paula Klug and Hannah Wahle for help with piloting the experiment. We thank Yue Hu for insightful discussion on earlier versions of this paper. We thank Syngyoo Choi and Shachar Kariv for generously providing their data and analysis code, as well as helpful pointers to other datasets. We thank for Matthew Polisson, John Quah, and Ludovic Renou for compiling and publishing a wide range of datasets. We thank the four anonymous reviewers for critical comments and comments that substantially increased the quality of this paper. The research was supported by a grant from the Deutsche Forschungsgemeinschaft (Bonn, Germany) to T.K. (grant DFG-KA 2675/4-3). Study 2 was preregistered (https://osf.io/wfd4z/).

[‡‡]Choi et al. and Kurtz-David et al. use a slightly different definition of the CCEI. In addition to the relaxation of definition 3, they apply a similar relaxation to definition 1. Further, Kurtz-David et al. [27], report 1 - CCEI. However, all nuances yield strongly related CCEI values. Our results hold independent of which definition of the CCEI is used (*SI Appendix*).

1. S. N. Afriat, On a system of inequalities in demand analysis: An extension of the classical method. *Int. Econ. Rev.* **14**, 460–472 (1973).
2. H. S. Houthakker, Revealed preference and the utility function. *Economica* **17**, 159–174 (1950).
3. P. A. Samuelson, A note on the pure theory of consumer's behaviour. *Economica* **5**, 61–71 (1938).
4. H. R. Varian, The nonparametric approach to demand analysis. *Econometrica* **50**, 945–973 (1982).
5. J. Rieskamp, J. R. Busemeyer, B. A. Mellers, Extending the bounds of rationality: Evidence and theories of preferential choice. *J. Econ. Lit.* **44**, 631–661 (2006).
6. D. W. Harless, C. F. Camerer, The predictive utility of generalized expected utility theories. *Econometrica* **62**, 1251 (1994).
7. A. Tversky, A critique of expected utility theory: Descriptive and normative considerations. *Erkenntnis* **9**, 163–173 (1975).
8. F. J. Nitsch, T. Kalenscher, Keeping a cool head at all times. What determines choice consistency? PsyArXiv [Preprint] (2020). https://doi.org/10.31234/osf.io/etyhx (Accessed 19 February 2020).
9. F. J. Nitsch, T. Kalenscher, How robust is rational choice? PsyArXiv [Preprint] (2021) https://doi.org/10.31234/osf.io/zv2m8 (Accessed 1 September 2021).
10. H. R. Varian, "Revealed preference" in *Samuelsonian Economics and the Twenty-First Century* (Oxford University Press, 2006), pp. 99–115.
11. S. Choi, S. Kariv, W. Müller, D. Silverman, Who is (more) rational? *Am. Econ. Rev.* **104**, 1518–1550 (2014).
12. H. B. Kim, S. Choi, B. Kim, C. Pop-Eleches, The role of education interventions in improving economic rationality. *Science* **362**, 83–86 (2018).
13. C. Hedge, G. Powell, P. Sumner, The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186 (2018).
14. R. Brouwer, I. Logar, O. Sheremet, Choice consistency and preference stability in test-retests of discrete choice experiment and open-ended willingness to pay elicitation formats. *Environ. Resour. Econ.* **68**, 729–751 (2017).
15. M. Mattmann, I. Logar, R. Brouwer, Choice certainty, consistency, and monotonicity in discrete choice experiments. *J. Environ. Econ. Policy* **8**, 109–127 (2019).
16. P. A. Samuelson, Consumption theorems in terms of overcompensation rather than indifference comparisons. *Economica* **20**, 1 (1953).
17. S. N. Afriat, Efficiency estimation of production functions. *Int. Econ. Rev.* **13**, 568–598 (1972).
18. J. Heufer, P. Hjertstrand, Consistent subsets: Computationally feasible methods to compute the Houtman–Maks-index. *Econ. Lett.* **128**, 87–89 (2015).
19. M. Houtman, J. Maks, Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve Methoden* **19**, 89–104 (1985).
20. R. Allen, J. Rehbeck, Measuring rationality: Percentages vs expenditures. *Theory Decis.* **91**, 265–277 (2021).
21. T. K. Koo, M. Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163 (2016).
22. J. Andreoni, J. Miller, Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* **70**, 737–753 (2002).
23. S. G. Bronars, The power of nonparametric tests of preference maximization. *Econometrica* **55**, 693–698 (1987).
24. P. Mayring, "Qualitative content analysis" in *A Companion to Qualitative Research* (Sage, 2004), vol. **1**, **pp.** 159–176
25. J. N. Rouder, J. M. Haaf, A psychometrics of individual differences in experimental tasks. *Psychon. Bull. Rev.* **26**, 452–467 (2019).
26. S. Choi, R. Fisman, D. Gale, S. Kariv, Consistency and heterogeneity of individual behavior under uncertainty. *Am. Econ. Rev.* **97**, 1921–1938 (2007).
27. V. Kurtz-David, D. Persitz, R. Webb, D. J. Levy, The neural computation of inconsistent choice behavior. *Nat. Commun.* **10**, 1583 (2019).
28. F. J. Nitsch, M. Sellitto, T. Kalenscher, Trier social stress test and food-choice: Behavioral, self-report & hormonal data. *Data Brief* **37**, 107245 (2021).
29. F. J. Nitsch, M. Sellitto, T. Kalenscher, The effects of acute and chronic stress on choice consistency. *Psychoneuroendocrinology* **131**, 105289 (2021).
30. D. Ahn, S. Choi, D. Gale, S. Kariv, Estimating ambiguity aversion in a portfolio choice experiment: Estimating ambiguity aversion. *Quant. Econom.* **5**, 195–223 (2014).
31. Z. Breig, P. Feldman, Revealing risky mistakes through revisions. *SSRN* (2021). https://doi.org/10.2139/ssrn.3975829 (Accessed 29 March 2022).
32. H. Quan, W. J. Shih, Assessing reproducibility by the within-subject coefficient of variation with random effects models. *Biometrics* **52**, 1195–1203 (1996).
33. A. Z. Enkavi *et al.*, Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5472–5477 (2019).
34. A. Pedroni *et al.*, The risk elicitation puzzle. *Nat. Hum. Behav.* **1**, 803–809 (2017).
35. T. Rakow, N. Y. Cheung, C. Restelli, Losing my loss aversion: The effects of current and past environment on the relative sensitivity to losses and gains. *Psychon. Bull. Rev.* **27**, 1333–1340 (2020).
36. L. R. Bruder, L. Scharer, J. Peters, Reliability assessment of temporal discounting measures in virtual reality environments. *Sci. Rep.* **11**, 7015 (2021).
37. K. N. Kirby, One-year temporal stability of delay-discount rates. *Psychon. Bull. Rev.* **16**, 457–462 (2009).
38. V. H. Aguiar, N. Kashaev, Stochastic revealed preferences with measurement error. *Rev. Econ. Stud.* **88**, 2042–2093 (2021).
39. J. K. Flake, E. I. Fried, Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Adv. Methods Pract. Psychol. Sci.* **3**, 456–465 (2020).
40. H.-K. Chung, A. Tymula, P. Glimcher, The reduction of ventrolateral prefrontal cortex grey matter volume correlates with loss of economic rationality in aging. *J. Neurosci.* **37**, 12068–12077 (2017).
41. D. R. Burghart, P. W. Glimcher, S. C. Lazzaro, An expected utility maximizer walks into a bar … . *J. Risk Uncertain.* **46**, 215–246 (2013).
42. D. Cohen *et al.*, Bounded rationality in C. elegans is explained by circuit-specific normalization in chemosensory pathways. *Nat. Commun.* **10**, 3692 (2019).
43. N. Haines *et al.*, Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox. PsyArXiv [Preprint] (2020). https://doi.org/10.31234/osf.io/xr7y3 (Accessed 9 November 2021).
44. V. M. Brown, J. Chen, C. M. Gillan, R. B. Price, Improving the reliability of computational analyses: Model-based planning and its relationship with compulsivity. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **5**, 601–609 (2020).
45. M. Brand *et al.*, Decision-making deficits of Korsakoff patients in a new gambling task with explicit rules: Associations with executive functions. *Neuropsychology* **19**, 267–277 (2005).
46. M. Brand, J. Schiebener, M.-T. Pertl, M. Delazer, Know the risk, take the win: How executive functions and probability processing influence advantageous decision making under risk conditions. *J. Clin. Exp. Neuropsychol.* **36**, 914–929 (2014).
47. S. V. Burks, J. P. Carpenter, L. Goette, A. Rustichini, Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 7745–7750 (2009).
48. S. Frederick, Cognitive reflection and decision making. *J. Econ. Perspect.* **19**, 25–42 (2005).
49. B. Gathmann, M. Pawlikowski, T. Schöler, M. Brand, Performing a secondary executive task with affective stimuli interferes with decision making under risk conditions. *Cogn. Process.* **15**, 113–126 (2014).

50. A. Tymula, L. A. Rosenberg Belmaker, L. Ruderman, P. W. Glimcher, I. Levy, Like cognitive function, decision making across the life span shows profound age-related changes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17143–17148 (2013).
51. Nitsch, Felix Jan, A Stress-Test of Economic Rationality, Doctoral Thesis, Heinrich-Heine-University Düsseldorf, (2021)
52. S. Palan, C. Schitter, Prolific.ac–A subject pool for online experiments. *J. Behav. Exp. Finance* **17**, 22–27 (2018).
53. E. Peer, L. Brandimarte, S. Samat, A. Acquisti, Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *J. Exp. Soc. Psychol.* **70**, 153–163 (2017).
54. J. R. de Leeuw, B. A. Motz, Psychophysics in a web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behav. Res. Methods* **48**, 1–12 (2016).
55. M. Castillo, D. L. Dickinson, R. Petrie, Sleepiness, choice consistency, and risk preferences. *Theory Decis.* **82**, 41–73 (2017).
56. E. Cettolin, P. S. Dalton, W. J. Kop, W. Zhang, Cortisol meets GARP: The effect of stress on economic rationality. *Exp. Econ.* **23**, 554–574 (2019).
57. A. Chakraborty, E. M. Calford, G. Fenig, Y. Halevy, External and internal consistency of choices made in convex time budgets. *Exp. Econ.* **20**, 687–706 (2017).
58. W. T. Harbaugh, K. Krause, T. R. Berry, GARP for kids: On the development of rational choice behavior. *Am. Econ. Rev.* **91**, 1539–1545 (2001).
59. G. Bedi, D. R. Burghart, Is utility maximization compromised by acute intoxication with THC or MDMA? *Econ. Lett.* **171**, 128–132 (2018).
60. M. Garagnani, The predictive power of risk elicitation tasks *SSRN* (2020). dx.doi.org/10.2139/ssrn.3692455 (Accessed 7 July 2022).
61. D. Lakens, Sample size justification. PsyArXiv [Preprint] (2021). https://doi.org/10.31234/osf.io/9d3yf (Accessed 25 January 2022).
62. S. Parsons, A.-W. Kruijt, E. Fox, Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Adv. Meth. Pract. Psychol. Sci.* **2**, 378–395 (2019).
63. C. L. Kimberlin, A. G. Winterstein, Validity and reliability of measurement instruments used in research. *Am. J. Health. Syst. Phar.* **65**, 2276–2284 (2008).
64. M. Polisson, J. K.-H. Quah, L. Renou, Revealed preferences over risk and uncertainty. *Am. Econ. Rev.* **110**, 1782–1820 (2020).