*****Please note that this manuscript has been accepted for publication in *Experimental and Clinical Psychopharmacology*. © 2021, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/pha0000545**

**A Multisample Demonstration of Using the Prolific Platform
for Repeated Assessment and Psychometric Substance Use Research**

Kasey Stanton***
University of Wyoming

Ryan W. Carpenter***
University of Missouri-St. Louis

Melissa Nance
University of Missouri-St. Louis

Taylor Sturgeon
Virginia Tech

Mara Villalongo Andino
Virginia Tech

Author Note: *** = Signifies joint first-authorship due to these authors making equal contributions in preparing this manuscript for publication. Correspondence concerning this paper should be addressed to Kasey Stanton at kaseystanton@gmail.com. Research ethics committee approval was obtained for this research, and all individual research participants provided informed consent for their participation. There is no funding to report for this research. The data from the first sample of participants (i.e., the Prolific daily diary study) have not been presented previously. Data from select measures from the other two samples (i.e., data from Prolific participants from two assessments approximately 2 weeks apart; cross-sectional data from undergraduates) have been published in one previous manuscript (see for more information: <https://osf.io/bcavr/>). However, all analyses presented are original and the scope of the current manuscript is distinct from any prior research. Data presented in the manuscript and data analytic syntax are available on the Open Science Framework website: <https://osf.io/x37ej/>

Abstract

The Prolific platform offers a potentially useful and efficient crowdsourcing option for repeated assessment substance use research, including for psychometric research requiring large samples. We present both (a) a series of practical recommendations for using Prolific and (b) data from multiple samples demonstrating Prolific's potential for efficiently collected repeated measures data. First, we present data from a 5-day daily diary protocol. We recruited a large sample ($N = 321$ at Day 1) screened for a history of self-identified mental health issues and weekly alcohol use. Participant adherence was good (82%) even without in-person contact. Alcohol use patterns conformed to theoretical expectations: Participants were more likely to drink on Fridays and Saturdays than other days, men drank more than women, and higher Alcohol Use Disorders Identification Test (AUDIT; Saunders et al., 1993) scores were associated with an increased likelihood of use and more overall drinking on a given day. Second, we present data from 429 Prolific participants screened for a history of mental health issues who completed assessments 2 weeks apart with strong retention ($N = 377$; 88%). We compare these data with data from undergraduates ($N = 529$) to demonstrate Prolific's utility for conducting psychometrically-oriented substance use research. Internal consistency estimates for measures from the Prolific data matched or exceeded those from the undergraduate data. Furthermore, measure scores showed strong temporal stability, and factor structures (e.g., AUDIT item-level structures) conformed to theoretical expectations. Collectively, these findings indicate that Prolific can be used successfully for repeated measures data collection.

KEYWORDS: Crowdsourcing; repeated measures designs; substance use; psychopathology; psychometrics

Public Significance Statement: We provide recommendations and data analytic examples showing how the online crowdsourcing Prolific platform can be used to efficiently collect longitudinal data from large samples of participants varying in their substance use and mental health histories. Our demonstrations indicate that Prolific can be used successfully for (a) studying how factors such as mood and interpersonal difficulties are associated with substance use over time and (b) improving how researchers assess substance use and mental health issues.

A Multisample Demonstration of Using the Prolific Platform for Repeated Assessment and Psychometric Substance Use Research

Crowdsourcing platforms such as Amazon Mechanical Turk (AMT) and Prolific have become widely used in psychological research over the past decade (Arditte et al., 2016; Paolacci & Chandler, 2014). Although these platforms have limitations, they can be useful for quickly collecting substance use and mental health data (e.g., Bounoua et al., 2021; Stanton & Watson, 2016; Strickland & Victor, 2020). Furthermore, crowdsourcing data collection methods likely will continue to play a prominent role in substance use research given the ongoing impact of the COVID-19 pandemic, as well as the unfortunate possibility of future pandemics (Tabish, 2020).

We build upon existing crowdsourcing research by demonstrating how the Prolific platform can be leveraged for brief longitudinal data collection. Intensive longitudinal designs increasingly are used in substance use research (e.g., Carpenter et al., 2019; Roos et al., 2020; Weiss et al., 2017), but with some exceptions (e.g., Strickland & Stoops, 2018; 2019), prior studies using crowdsourcing platforms have relied on cross-sectional designs. Furthermore, prior reviews describing the use of crowdsourcing methods in substance use research have focused primarily on AMT (e.g., Arditte et al., 2020; Mellis & Bickel, 2020; Schick et al., 2019), such that demonstrating the utility of other platforms such as Prolific may be informative for researchers.

We present descriptive information and example analyses from two Prolific-recruited samples to demonstrate how this platform can be used to conduct repeated measures substance use research.¹ First, we present data from a sample of 321 Prolific participants who completed a 5-day daily diary protocol targeting substance use and mood assessment with good participant

¹ More detailed descriptions of both of these Prolific samples and of the study goals associated with them are provided in Online Appendix S1.

adherence across the study protocol. We examine patterns of alcohol use, other substance use, mood, and interpersonal functioning over the 5-day study period to determine the extent to which Prolific can be used to collect data confirming to theoretical expectations. For example, demonstrating that more participants reported drinking and higher levels of drinking on Fridays and Saturdays compared to other days would provide basic evidence of Prolific's utility, as would finding that participants with more self-reported alcohol problems at baseline drink more across the 5-day study period. Other examples of theoretically consistent patterns of results would include participants reporting higher average levels of positive mood and lower levels of negative mood on weekend days and Fridays compared to other days of the week (Stone et al., 2012).

We also present data and demonstrative analyses from second a Prolific-recruited sample of 429 participants who completed substance use and psychopathology measures at two time points roughly 2 weeks apart. These second series of demonstrations focus specifically on longitudinal research questions relevant to psychometrics, assessment, and measure development. They illustrate how Prolific can be used to recruit large samples needed for psychometrically-oriented data analytic procedures such as factor analysis (Sellbom & Tellegen, 2019; Strickland & Stoops, 2019), as well as for conducting longitudinal analyses that are important but often neglected in the measure development process (e.g., examinations of test-retest reliability; Clark & Watson, 2019). We also compare data from this second Prolific sample with cross-sectional data from 529 undergraduates who completed the same measures as did this sample of Prolific participants. For example, we compare the amount of missing data for the undergraduate sample with this second Prolific sample, and we also evaluate the consistency

of psychometrically-oriented analyses across samples (e.g., are factor structures similar across the undergraduate and Prolific samples).

Practical Considerations for Participant Recruitment and Longitudinal Data Collection

Prior to presenting sample descriptions and illustrative analyses, we describe strategies for targeting participant recruitment for longitudinal substance use and mental health research. First, regarding participant compensation, Prolific enforces a minimum \$6.50 USD per hour compensation rate (see <https://tinyurl.com/5bs3sckj>), which is important given previously documented issues concerning fair compensation in research using AMT (see Mellis & Bickel, 2020 for discussion relevant to substance use research). Researchers can compensate participants at higher rates, but Prolific automatically pauses studies if hourly compensation rates fall below this required minimum.

Regarding participant recruitment, Prolific requires participants to verify their identities (this information is not provided to researchers to ensure participants' anonymity). Participants provide their first and last names, date of birth, country of residence, a photo of themselves, and a picture of an ID for verification (<https://tinyurl.com/34yrtafb>). Upon verification, the Prolific system assigns each participant a unique ID that researchers can request when administering measures via Qualtrics or other survey platforms to link data across timepoints.

A limitation of crowdsourcing platforms is the inability to have in-person contact with participants. However, researchers can interview participants using Zoom or similar videoconferencing platforms (see <https://tinyurl.com/kt7nbtpk>), though we are not aware of prior research using this feature. When using survey-based research designs, researchers can directly message participants via the Prolific platform and send reminders to improve participant retention in longitudinal studies. Although Prolific does not have a standalone smartphone app,

Prolific can be accessed via a web browser on smartphones. Messages sent by researchers through the Prolific platform also are sent to participants' registered email addresses.

Finally, participant recruitment can be targeted using screening filters that require no additional cost, with a few exceptions (e.g., recruiting a nationally representative sample). Screening filter information is self-reported by participants (e.g., self-reported histories of mental health disorders are not corroborated by mental health professionals). Participants can be filtered based on demographics (e.g., see <https://tinyurl.com/4unuvnb3> for recommendations provided by Prolific for ensuring gender and other demographic representation in samples), as well as their current levels of use of alcohol or other substances (e.g., smoking more than 5 times per day). Other filters (e.g., experiences of chronic pain) also may be of interest when studying substance use among specific populations (e.g., Carpenter et al., 2019; Schick et al., 2019), and multiple filters can be used in tandem (e.g., using filters for both alcohol and tobacco use).

Additionally, multiple filters for screening participants based on mental health history are available, including: (a) currently taking antidepressant, antianxiety, or antipsychotic medication (or some combination thereof), (b) currently or previously having a diagnosed mental health condition (any self-identified disorder), and (c) experiencing functional impairment related to having mental health issues. However, researchers cannot screen participants for specific mental health disorders other than filtering participants based on having a self-identified history of autism spectrum disorder or a neurocognitive disorder.

Prolific Sample 1: Daily Diary Study

Sample

All participants in this and the subsequent samples provided informed consent, and all research procedures received institutional review board (IRB) approval. Approval from the IRB

was obtained to share de-identified data, which can be found here: <https://osf.io/x37ej/> (Stanton, 2021). Study analyses were not pre-registered. All samples described subsequently had more than 300 participants at initial assessments to ensure analyses were well-powered. We provide descriptions of data exclusions (i.e., measures that were administered to participants but not included in analyses) and missing data following a review of the study sample characteristics.

All daily diary study participants were 18 years or older. The goal of this broader study was to investigate day-level associations for substance use with interpersonal and mental health experiences. Therefore, to increase the likelihood of sufficient variation in these constructs, participants were screened using filters for (a) self-reported history of any mental health disorder diagnosis (i.e., “do you have or have you had a diagnosed mental illness”) and (b) drinking alcohol at least once weekly on average.

Initially, 340 Prolific participants enrolled in the 5-day daily diary study occurring during the Fall 2020. Fifteen participants completed few to no baseline items on Day 1 and were excluded. Four additional participants were also excluded: One for entering an incorrect ID code; one for failing an obvious attention check at the beginning of the Day 1 survey (please type the word “start”); and two for failing an obvious validity check item (e.g., select “strongly agree”) on two or more daily surveys. Thus, only a small percentage (5.6%; 19 of 340) were removed for invalid or non-responding and the final analyzed sample consisted of 321 participants.

Demographics are presented in Table 1. To summarize this information briefly, most participants identified as White or European American, with more variation for other demographic variables. Demographics were similar to crowdsourced samples described in reviews of studies using AMT in some ways (e.g., Arditte et al., 2016; Paolacci & Chandler,

2014), including participants tending to be relatively young and a substantial portion of them being unemployed.

Study Procedure and Adherence

On Day 1, participants completed a baseline assessment consisting of personality and substance measures. Immediately following this, they completed a brief daily assessment. Collectively, these two series of measures took participants roughly 44 minutes on average to complete, and participants were compensated with \$10 USD. This compensation amount and all amounts discussed subsequently were determined by our research team and exceeded Prolific's recommendations for participant compensation. Researchers have the option of providing participants with "bonus" compensation, but we did not use any bonus compensation incentives in either of our studies. Prolific provides an approximate estimate of the average participant compensation per hour after removing outliers in completion time, which was estimated to be approximately \$15.60 for all Day 1 measures administered.

Participants then completed the same daily survey over the next 4 days, resulting in five total daily data collection points. Participants were compensated with \$3 USD for each of the Day 2-5 surveys, which were expected to require 10 minutes on average. Prolific's estimated hourly compensation rate for these daily surveys after removing outliers was nearly \$20 per hour on average.

We completed several waves of data collection (with 75-130 participants per wave) so that we could monitor and correct administration and adherence issues if needed. Across waves, Day 1 was always on a Wednesday and Day 5 was always on a Sunday to ensure all participants reported over several weekdays and a weekend. Participants were instructed to complete daily assessments as close to bedtime as possible so that their daily ratings would reflect the entirety of

the day. In the first recruitment wave, the survey became available at 4pm Eastern Time; this was changed to 7pm Eastern Time for subsequent recruitment waves to ensure participants did not complete assessments too early. For each assessment, participants received notifications through the Prolific platform that directed them to Qualtrics surveys, which were formatted to be mobile-friendly if participants accessed surveys via a smartphone. Participants were sent a reminder to complete daily surveys each day and could contact research staff with questions via the Prolific platform, but there was no researcher-participant contact otherwise.

Participants completed 1,366 daily surveys. Most were completed between 6pm and 9pm Eastern Time ($n = 1,112$; 81.4%). Across all recruitment waves, surveys remained open until the late morning or early afternoon of the following day (e.g., Wednesday surveys remained open until the early afternoon on Thursday), although only a small number of surveys were completed after 6am the following day ($n = 43$; 3.1%). For these 43 surveys, it was unclear whether participants were answering questions about their current day or the previous day, so we dropped them from analyses. We kept surveys that were completed between 12am and 6am ($n = 84$; 6.2%) because these may have been completed prior to going to sleep after a late night of substance use. Participation declined somewhat across the 5-day protocol, with 320 included responses on Day 1 and 245 on Day 5. The final dataset included surveys from 1,320 days out of a possible 1,605 (321×5) for an adherence of 82.2%. Of these, 2 days were only partially completed, but were included in analyses.

Measures

Baseline Substance Use

Participants in this sample and others described subsequently completed measure batteries that were too extensive to integrate in their entirety here and included measures that were not

central to our study aims (e.g., retrospective ratings of childhood psychopathology). As a result, we focused on using data from measures most relevant to our demonstrative analyses (e.g., examination of alcohol use patterns). More detailed description of the full study battery and data exclusions are provided in Online Appendix S1.

First, participants completed the 10-item Alcohol Use Disorders Identification Test (AUDIT; Saunders et al., 1993), which assesses alcohol use and associated problems. Participants responded to the AUDIT items in reference to the past 6 months. Scores were classified into Low (< 8), Hazardous (8-15), Very Hazardous (16-19), and Severe (> 19) categories. Participants also completed the 8-item Cannabis Use Disorders Identification Test-Revised (CUDIT-R; Adamson et al., 2010) to assess average level of cannabis use and its potential impacts. Scores were classified into Low (< 8), Problems (8-11), and Probable Use Disorder (> 11) categories. For both measures, the response scale varies according to the item in question (e.g., *Never* to *4+ Times per Week* for level of cannabis use; *Never* to *Daily* when asking about specific consequences of use).

Daily Affect

At each daily assessment, positive affect, negative affect, and fatigue were assessed using select items from the Positive and Negative Affect Schedule-Expanded Form (Watson & Clark, 1999). Participants were asked to rate their affective experiences for that day on a 5-point Likert-type scale ranging from 1 (*very slightly or not at all*) to 5 (*extremely*) for each item administered. Participants completed positive affect items assessing joviality (joyful, excited, cheerful), self-assurance (fearless, bold, strong), and attentiveness (attentive, determined, alert), as well as negative affect items assessing anger (angry, hostile, irritable), sadness (sad, blue, lonely), fear (nervous, afraid, jittery), and guilt (guilty, ashamed, disgusted with self). Item

scores were summed to create composite Positive (9 items) and Negative Affect (12 items) scales. We also assessed Fatigue using 3 items (tired, sleepy, sluggish).

We assessed reliability of these indicators using a generalizability theory approach (Cranford et al. 2006; Shrout & Lane, 2012). Specifically, we examined reliability of our indicators in four ways: The ability to differentiate persons on a single fixed day (R_{1F}), the ability to differentiate between persons on different days (treating day as random; R_{1R}), the ability to differentiate based on their responses over all study days (R_{KF}), and the ability to detect systematic within-person changes from day to day (R_C). R_{1F} was high for all three indicators (Positive Affect = .89; Negative Affect = .92; Fatigue = .87). R_{1R} was moderate (Positive Affect = .62; Negative Affect = .66; fatigue = .58). R_{KF} was excellent (Positive Affect = .97; Negative Affect = .98; Fatigue = .97). R_C was good (Positive Affect = .79; Negative Affect = .82; Fatigue = .77). Thus, measures generally evidenced good to excellent reliability. The ability to differentiate persons across different days (R_{1R}) was somewhat lower, but this is not unusual and is largely not an issue in the current study because all participants started on a Wednesday and ended the study on a Sunday. R_{1R} would be more relevant, for example, if we only had weekend data for some participants and only weekday data for others.

Daily Alcohol and Other Substance Use

Each day, participants reported the number of standard drinks consumed (a) in the current day and (b) the previous day. Definitions of a standard drink, with a visual aid, were provided to participants within the survey. Participants also reported on substance use for the current day and the previous day by indicating which substances they consumed from a list of options (e.g., nonprescribed cannabis, nonprescribed opioid use, other substances such as amphetamines). Asking participants about the prior day allowed us to assess whether additional drinks or

substances were consumed during the prior day after the daily report was completed, acknowledging that participants may be less accurate in recalling previous day experiences.

Daily Negative Interpersonal Experiences

Participants were administered 14 items describing experiences of interpersonal conflict and rejection (e.g., “got into an argument;” “rejected by someone else”). Each day, they were asked to indicate all experiences that occurred that day. These dichotomous responses were summed to create an indicator of daily negative interpersonal experiences.

Prolific Sample 1 Results

Descriptive Information

As anticipated, due to screening Prolific participants based on having a significant mental health history, most participants reported current psychotropic medication use (53.6%) and a sizable portion of participants reported current psychotherapy (38.3%). A minority (5.9%) reported current substance use treatment. The mean AUDIT score was 7.6 ($SD = 7.1$; $\omega = .89$). Most participants were classified by the AUDIT as Low (66.4%), and the remaining third scored as Hazardous (18.7%), Very Hazardous (5.6%), or Severe (9.3%). The mean CUDIT-R score was 4.4 ($SD = 6.7$; $\omega = .89$). Most participants were classified as Low (77.9%), with the remaining classified as having Problems (6.8%) and Probable Use Disorder (15.3%).

Daily Substance Use

We focus on the extent to which participants reported alcohol use in an expected manner to provide researchers with a sense of the nature of daily diary data that can be obtained using Prolific. Participants reported alcohol use on 311 days (23.56%). Responses were relatively consistent for previous-day report, with participants reporting alcohol use on the previous day on 304 days (23.04%). Participants reported nonprescribed use of at least one other substance on

315 days (23.86%). Participants most often reported cannabis use ($N = 215$, 16.29%), followed by use of nonprescribed opioids ($N = 37$, 2.80%), nonprescribed sleep medications ($N = 30$, 2.27%), and other substances ($N = 71$, 5.38%).²

On drinking days, participants reported an average of 2.67 drinks ($SD = 2.01$). Participants reported more drinks when reporting on the previous day ($M = 3.58$, $SD = 2.90$), which may reflect that participants sometimes had additional drinks after completing a survey or, possibly, a bias to over-report drinking from the previous day. Participant reports of drinking today and the previous day (adjusted to match) were highly correlated, $r(1,318) = .71$. This reduced, but only slightly, when removing non-drinking days, $r(1,318) = .64$, suggesting that the association was not dependent on days without drinking (which may be easier to recall).

Figure 1 presents drinking behavior over the 5-day protocol (again, all participants began on a Wednesday). We present both the percentage of days with any drinking and the average number of drinks on drinking days. Drinking was most likely on Fridays (27.78%) and Saturdays (26.07%) and least likely on Wednesdays (19.06%) and Sundays (21.63%). On drink days, participants reported the most drinks on Fridays ($M = 2.83$, $SD = 2.00$), followed by Thursdays ($M = 2.77$, $SD = 2.07$). They reported the fewest drinks on Sundays ($M = 2.34$, $SD = 1.84$).

A multilevel logistic model with a person-level random intercept found that participants were more likely to drink on the weekend (Fridays and Saturdays) than on weekdays (OR = 1.42, 95% CI [1.05, 1.93], $p = .024$). However, in a multilevel linear regression model examining drink days, number of drinks did not differ between the weekend and weekdays ($b = 0.21$, 95% CI [-0.16, 0.57], $p = .271$). Men and women did not differ in likelihood of alcohol use (OR =

² The reports of individual substances sum to more than 315 because participants could report the use of multiple substances on the same day.

1.27, 95% CI [0.76, 2.11], $p = .355$), but, on drink days, men ($M = 3.16$, $SE = 0.23$) drank more than women ($M = 2.15$, $SE = 0.17$; $b = 1.01$, 95% CI [0.43, 1.58], $p < .001$). Higher AUDIT scores were associated with greater likelihood of drinking on any given day (OR = 1.14, 95% CI [1.11, 1.17], $p < .001$) and drinking more on drink days ($b = 0.12$, 95% CI [0.09, 0.15], $p < .001$).

Contextual Factors

We also present patterns in Figure 1 for several factors that researchers often examine in conjunction with substance use: Positive affect ($M = 2.07$, $SD = 0.74$), negative affect ($M = 1.92$, $SD = 0.77$), fatigue ($M = 2.88$, $SD = 1.14$), and daily negative interpersonal experiences ($M = 1.78$, $SD = 1.51$). These factors might be expected to follow a weekly pattern, such that participants report more positive affect and less negative affect and fatigue, and fewer negative experiences during the weekend compared to the rest of week. Such patterns are visible in Figure 1 and were confirmed by multilevel linear regression models: On weekends (Friday, Saturday, Sunday),³ relative to weekdays, participants reported more positive affect ($b = 0.06$, 95% CI [0.01, 0.11], $p = .032$), less negative affect ($b = -0.22$, 95% CI [-0.27, -0.17], $p < .001$), fatigue ($b = -0.39$, 95% CI [-0.47, -0.31], $p < .001$), and fewer negative interpersonal experiences ($b = -0.30$, 95% CI [-0.43, -0.17], $p < .001$).

Daily Diary Study Summary

Our findings demonstrate that Prolific can be used to successfully target participant recruitment (e.g., the AUDIT indicated that a sizeable subset of participants had possible issues with alcohol use) and retain participants across a brief daily diary protocol. In fact, adherence (82.2%) across the 5-day study period was above that observed in a meta-analysis of daily diary

³ For these analyses we classified Sunday as part of the weekend whereas we did not for alcohol use. This is because alcohol consumption primarily happens in the evening, and Sunday evening is typically the transition point back to the workweek for most individuals.

studies of substance use (76.8%; Jones et al., 2018). This strong adherence rate suggests that it may be possible to conduct longer or more intensive protocols using Prolific (e.g., multiple reports per day), although researchers may want to consider additional points of contact (e.g., more frequent reminders) or incentives (e.g., bonus compensation) to ensure adherence remains high. Even these additional steps, however, would require relatively little in terms of researcher effort and resources. Importantly, we collected a larger sample than is typical for intensive longitudinal work (Jones et al., 2018). In terms of power, this has advantages, as increasing the number of participants benefits power more than increasing the number of assessments, even holding person-assessments constant, especially when the observed frequency of substance use is less (or more) than 50% of observations (Lane & Hennes, 2019). However, it would be reasonable to recruit a smaller number of participants to complete a longer or more intensive protocol for an equivalent cost in compensation.

Results of our longitudinal analyses aligned with theoretical expectations regarding both (a) temporal patterns (e.g., drinking being most likely on Fridays and Saturdays) and (b) associations amongst indicators of substance use, mood, and interpersonal experiences. Collectively then, our results indicate that Prolific represents a potentially sound option for collecting longitudinal daily diary data of good quality, although it would be beneficial for future studies to examine Prolific's utility when using other study designs (e.g., multiple assessments per day).

Prolific Sample 2: Test-Retest Study

Samples

Prolific Participant Demographics and Procedure

Our second sample of Prolific participants ($N = 429$) were US adults (age 18+) who completed two assessments roughly 2 weeks apart in the Fall of 2020. Participants were

screened based on a history of mental health issues using the same filter described previously. These data were collected as part of a broader study focused on improving transdiagnostic mental health assessment (see Stanton et al., 2021). All analyses presented are original, and more information regarding this study's goals and data exclusions (i.e., scores from measures that were administered but not included here) are provided in Online Appendix S1.

Demographics are presented in Table 1 and were very similar to the daily diary study sample (e.g., regarding employment status). Both Prolific sample participants and the participants from the undergraduate sample described subsequently completed surveys administered via Qualtrics, which again included a mobile-friendly version for participants who chose to complete study measures in this manner.

We obtained data from an initial sample of 439 Prolific participants, but nine of these participants (a) failed to pass an obvious initial attention check (i.e., type the word "start") and/or (b) answered few to no study items. Participants also were administered two other obvious validity items (i.e., please answer "strongly agree"). One participant failed both validity items, and removal of this participant's data and data from other the nine cases described reduced our sample size to 429 participants. Notably then, only a very small percentage (2.3%; 10 of 439) of cases were removed due to invalid responding. There were very little missing item-level data after removing invalid cases (e.g., the AUDIT was administered, but there were only a handful of missing AUDIT item scores in this dataset). Data collection was efficient, as all data from the initial assessment were collected after the study had been posted for less than 2 days.

Most of the original 429 Prolific participants ($N = 377$; 87.8%) completed the second assessment occurring approximately 2 weeks after the first (13.7 days; $SD = 1.3$; mode of 13 days; range from 13-21 days). From hereon, we refer to the first and second assessments as

“Time 1” and “Time 2,” respectively. The original Time 2 sample size was 395, but 15 of these failed to respond to nearly all items or failed multiple validity checks (i.e., failing to select an instructed response). Additionally, three participants made errors when entering their Time 2 Prolific IDs at Time 2, which prevented us from linking their data and reduced the final Time 2 sample size to 377. Participants were compensated with \$7.50 USD for completing each survey assessment, and Prolific’s estimated average participant completion time after removing outliers for each survey was approximately 30 minutes. Therefore, estimated compensation was approximately \$15 per hour.

Of the 377 participants who completed the Time 2 assessment, 92% (347) completed this assessment within a 13- to 15-day window. This suggests that researchers may be able to obtain follow-up data from participants within precise timeframes, thereby avoiding complications introduced by participants varying widely in their timing of completing follow-up assessments (e.g., complications when evaluating retest coefficients if some participants complete measures after a week but others after 3-4 weeks). Furthermore, our retention rate of 87.8% was achieved by sending a single reminder to participants via the Prolific system at follow-up and compares favorably to other studies. For example, some studies using online administration with undergraduate samples report 2-week retention rates of 65-75% (Chmielewski et al., 2017; Stanton & Watson, 2020), with 2-week retention rates of roughly 80% reported for some studies including in-person administration (Jonason & Webster, 2010; Watson et al., 2019).

Undergraduate Comparison Sample Demographics and Procedure

As reviewed, we also present cross-sectional data from undergraduate participants. These participants were recruited from the psychology department subject pool at a large southern university over the duration of the Fall 2020 semester (early September through mid-December).

All participants received course credit for their participation. Participants were required to be 18 years or older but were not screened based on any other criteria. This sample initially consisted of 613 participants, but data from 84 participants were removed, reducing our final sample to 529 participants. Reasons for data removal included 71 participants failing to respond to nearly all study items as well as 13 participants failing the same attention checks used with our second Prolific sample. Sample demographic information for this sample is presented in Table 1. There was somewhat more variability in some demographic characteristics such as racial identity in the undergraduate than in the Prolific samples, though less variation in other demographic variables such as sexual orientation.

Measures

Substance Use and Mental Health History

Participants completed a range of items assessing their substance use and mental health histories (e.g., currently receiving psychotherapy). We also asked participants to self-identify whether they had ever been diagnosed with a range of mental health disorders. Although participants may lack specific knowledge of various disorder criteria, we still present these data to provide a general sense of the types of mental health issues Prolific participants oversampled for psychopathology history may report.

Substance Use Questionnaires

The 10-item AUDIT and 8-item CUDIT-R were administered to assess alcohol and cannabis use in reference to the past 12 months. Additionally, participants completed an 8-item version of the Drug Use Survey (DUS; Clark & Watson, 1999) to assess lifetime use of other substances (e.g., amphetamines, tranquilizers). The DUS asks participants to report the number

of times they have used specific substances on a 7-point scale ranging from 0 (*Never*) to 6 (*40 times or more*), and ratings for all DUS items were summed to create a composite score.

Personality Pathology and Symptom Measures

Participants completed select scales from the Comprehensive Assessment of Traits Relevant to Personality Disorder (CAT-PD; Wright & Simms, 2014). Given that we administered many different CAT-PD scales, we summed scores from scales that were strongly correlated to create composite variables to model trait domains from *DSM-5* Section III Alternative Model for Personality Disorders (AMPD). All scales summed to create composites were correlated $\geq .50$ in the undergraduate data and Prolific data across timepoints. Participants responded to all items using a 5-point scale ranging from 0 (*very untrue of me*) to 4 (*very true of me*). Composites included a Disinhibition composite created by summing CAT-PD Irresponsibility (7 items) and Non-Perseverance (6 items) scores; an Antagonism composite representing a sum of Callousness (7 items) and Manipulativeness (6 items); Anxiousness (7 items) and Depressiveness (6 items) summed to create a Negative Affectivity composite; and Unusual Beliefs (7 items) and Unusual Experiences (7 items) used to create a Psychoticism composite. Scores from the CAT-PD Social Withdrawal scale also were used as an indicator of the AMPD Detachment domain.

Additionally, participants completed the Mood Disorder Questionnaire (MDQ; Hirschfeld et al., 2000), which assesses a lifetime history of hypomania/mania symptoms using 13 items. Participants responded to these items using a dichotomous “Yes” or “No” format. We summed scores on these 13 items to create a dimensional hypomania history score. We also report the percentage of participants who screened positively for a possible hypomania history. A positive MDQ screen requires endorsement of seven or more MDQ items. It also requires

participants to endorse that (a) multiple symptom experiences occurred during the same time period, and (b) symptoms experienced caused moderate to severe impairment. Finally, participants completed the 10-item Dysphoria scale (e.g., “felt discouraged”) from the Expanded Version of the Inventory of Depression and Anxiety Symptoms (IDAS-II; Watson et al., 2012). These 10 items were summed to create a dimensional Dysphoria scale score, and we also report the number of participants who exceeded a Dysphoria screening score cutoff of 28. Scores exceeding this cutoff indicate an increased likelihood of meeting criteria for one or more internalizing disorders (e.g., major depressive disorder, generalized anxiety disorder; Stasik-O’Brien et al., 2019).

Prolific Sample 2 Results

Descriptive Statistics

Information regarding participants’ mental health and substance use histories are provided in Table 2. As anticipated, due to screening Prolific participants based on having a significant mental health history, most participants reported current psychotropic medication use (52.9% at Time 1), and roughly a third reported current psychotherapy (32.4% at Time 1).

The Prolific and undergraduate samples differed markedly in terms of mental health history. This was expected given that Prolific participants were screened on psychopathology history whereas undergraduate participants were not. However, this provides evidence that the Prolific screening filters worked as intended. For example, current antidepressant use was over three times higher in the Prolific (42% and 41.4% at Times 1 and 2, respectively) than in the undergraduate sample (13%), and nearly three-fourths (> 70% across timepoints) of Prolific participants self-identified as having been diagnosed with a depressive disorder. Although this Prolific sample was not screened for substance use, 10.3% of Time 1 participants reported

having ever sought treatment for substance use (versus 2.1% of the undergraduate sample), and 9.3% of Time 1 participants reported having been diagnosed with any alcohol or other substance use disorder (versus 1.5% of the undergraduate sample; Time 2 frequencies generally were slightly lower for these and other variables; see Table 2).

Table 3 provides screening measure frequencies, measure means and *SDs*, and internal consistency estimates provided as coefficient omega values. For nearly all measures reported, coefficient omega values from the Prolific sample were equal to or exceeded those from the undergraduate sample (e.g., $> .90$ across timepoints for the AUDIT in the Prolific data; $.85$ in the undergraduate data).

We computed Cohen's *d* effect size differences to compare measure scores for the Time 1 Prolific data with the undergraduate data, which also are provided in Table 3. Mean scores for most psychopathology measures were higher for the Prolific than undergraduate sample, with sizeable effect size differences in some cases (e.g., $d = .83$ for Detachment). Lifetime substance use history scores assessed via the DUS were higher in the Prolific than in the undergraduate data ($d = .78$), but CUDIT-R and AUDIT differences were trivial (both $ds < .14$).

When considering AUDIT and CUDIT-R classification levels across Time 1 and undergraduate data, more Prolific participants (i.e., 76.7%) than undergraduates (67.6%) were classified as Low by the AUDIT; however, a greater percentage of Prolific participants (5.6% at Time 1 versus 1.9% of undergraduates) were classified in the AUDIT Severe range. Regarding CUDIT-R classification, slightly higher percentages of Prolific participants (e.g., 10.3% at Time 1 for Probable Use Disorder) were classified as having Problems or Probable Use Disorder than in the undergraduate sample (7.8% Probable Use Disorder).

Substance Use Item Factor Analyses

Factor analysis commonly is used in psychometric research (Sellbom & Tellegen, 2019), and we present a series of factor analyses demonstrating Prolific’s utility for conducting psychometrically-oriented longitudinal research. Many different types of factor analyses could be conducted using the data across the samples and timepoints available here. Acknowledging this, we focus on basic demonstrative analyses examining single-factor structures of item scores from the AUDIT, CUDIT-R, and DUS (e.g., examining a single-factor structure of the 10 AUDIT items). Although these measures are widely used and validated, these analyses show that all item scores from each measure reflect a common factor as would be expected theoretically (however, see Watts et al., 2021 for discussion regarding the unidimensionality of substance use item sets and the substance use disorder criteria).

All factor analyses were exploratory factor analyses (EFA) conducted using principal axis factoring with squared multiple correlations as the communality estimates. All EFAs were based on polychoric correlation matrices given our focus on analyzing item-level ratings. We also evaluated the similarity of EFA results across timepoints and samples using comparability coefficients. With this approach, coefficients can be computed using both sets of data. For example, when evaluating the similarity in single-factor structures across the two samples, one coefficient can be generated using the Prolific dataset for scoring and a second coefficient using the undergraduate dataset for scoring (Everett, 1983). These coefficients can then be averaged to obtain an estimate of overall comparability, and comparability coefficients $\geq .90$ indicate strong factor similarity (Everett, 1983).

All items from each measure (e.g., when factor analyzing the 10 AUDIT item ratings) loaded strongly ($\geq .50$) onto a single factor across samples and time points (i.e., the Time 1 and Time 2 Prolific data; the undergraduate data). Factor loadings are presented in Online

Supplemental Table S1. Across Time 1 and Time 2 in the Prolific sample, (a) all AUDIT items loaded $\geq .55$ on a common factor, (b) all CUDIT-R items loaded $> .80$ onto a common factor, and (c) all DUS items loaded $\geq .74$ onto a common factor. All comparability coefficients indicated very strong factor similarity for the AUDIT, CUDIT-R, and DUS factor structures across the Prolific sample timepoints (i.e., all coefficients were $\geq .98$). Thus, factor structures were consistent across Prolific timepoints and with theoretical expectations in all cases.

Cross-sample factor similarity also was strong when comparing the Prolific and undergraduate sample results. For example, factor structures of both the AUDIT and CUDIT-R items showed very strong similarity across the Time 1 Prolific data and the undergraduate data (all comparability coefficients $\geq .97$). Comparability across the Prolific Time 1 and undergraduate results was somewhat weaker for the single-factor structure of the DUS items when using the undergraduate data for coefficient scoring (coefficient = .89), although comparability (coefficient = .95) was stronger when using the Prolific Time 1 data for scoring. The average of these coefficients still was .92, suggesting that the single-factor structures of the DUS were comparable across the Prolific Time 1 data and undergraduate data overall.

Stability of Scores Across Timepoints

Finally, we demonstrate how Prolific can be used to obtain data needed for other psychometrically-oriented analyses involving examination of the test-retest stability of measure scores. Use of data from the 377 Prolific participants who completed both study assessments for computing test-retest correlations exceeds even stringent sample size recommendations (e.g., 300 participants; Watson, 2004), which is important to note given that researchers often use very small sample sizes for test-retest analyses if they are conducted at all (e.g., samples of 50-75

participants; Watson, 2004). Test-retest stability coefficients reported as Pearson correlations conformed with theoretical expectations when examining stability over the 2-week study period, as shown in Table 4. For example, the coefficients for the three substance use measures administered (i.e., the AUDIT, CUDIT-R, and DUS) were $\geq .90$, consistent with expectations that past year (AUDIT and CUDIT-R) and lifetime reports of substance use (DUS) should not change meaningfully over a 2-week retest timeframe. Temporal stability estimates for other measure scores (e.g., personality pathology) also aligned with theoretical expectations.

Test-Retest Study Summary

Results from this second Prolific sample provide further evidence that filters available via the Prolific platform are useful for efficient and targeted data collection. Retention across the 2-week assessment interval also was strong in the Prolific sample. The analyses based on this sample demonstrate how Prolific can be used to collect data from relatively large samples that are necessary, but often difficult to obtain, to adequately conduct longitudinal psychometric analyses. The Prolific data used for these analyses also compared favorably in many ways to data collected from the undergraduate sample (e.g., fewer missing data than in our undergraduate sample). This is noteworthy given that researchers may have concerns about participants recruited from crowdsourcing platforms responding carelessly to study items, although we acknowledge that concerns regarding response validity often apply to undergraduate samples as well. Taken together, these results suggest the Prolific platform has significant potential for use in psychometrically-oriented research.

General Discussion and Future Directions

Our data from two Prolific samples collected across seven assessments (i.e., 5 days of daily diary data collection in our first sample; an initial and follow-up assessment in our second

sample) indicate that Prolific can be used to collect data of reasonably high quality for repeated measures substance use research. The complementary series of analyses presented using data from the daily diary and test-retest studies demonstrate specific ways researchers can use Prolific to address contemporary research questions, and it was encouraging that results across samples conformed to theoretical expectations (e.g., the likelihood of alcohol use was increased on Fridays and Saturdays compared to other days for the daily diary study). Furthermore, data from both samples were collected with relatively little time and effort cost on the part of our research team. Prolific's built-in screening filters aided in ensuring an efficient data collection process, and descriptive information indicated that these screening filters were useful for targeting participant recruitment to address specific substance use and mental health research questions. Screening filters may not be as readily available or as straightforward to apply when recruiting participants from other convenience samples (e.g., undergraduate samples), and Prolific's process of verifying participant identities may be another advantage over other online recruitment platforms where doing so may be more difficult (e.g., Facebook).

Our examples are by no means exhaustive of how Prolific could be used by substance use researchers. For example, it would be interesting to evaluate Prolific's utility for conducting research using more intensive or extended sampling approaches. This could include examining the extent to which adherent, valid, and reliable responses can be obtained when assessing participants over longer time periods and/or multiple times per day (i.e., ecological momentary assessment [EMA]). Instead of using Qualtrics as we did here, Prolific potentially could be paired with software specifically designed to collect EMA data. Relatedly, retention was strong for our second Prolific sample over a 2-week period, and results were highly consistent across

assessments. However, it would be interesting to examine retention and patterns of results over more extended follow-up periods (e.g., 6 or 12 months).

Finally, although we found our results encouraging overall, we are not arguing that crowdsourcing platforms such as Prolific can or should replace in-person research efforts. For example, it is often useful to incorporate multimethod assessment when validating new measures (e.g., to examine convergence across self and interview ratings). However, it is difficult to determine the extent to which Prolific is useful for conducting multimethod research due to most existing research with this platform having focused on self-report. Furthermore, although crowdsourcing is efficient for collecting large samples, samples recruited with these methods may have their own idiosyncrasies (e.g., possible personality differences for crowdsourced samples compared with other sample types; Arditte et al., 2016; Paolacci & Chandler, 2014; Peer et al., 2017). Given the vagaries of the internet, Prolific may also be vulnerable to sudden and large shifts in its participant pool, both in terms of response quality and participant demographics (e.g., temporary increases in the number of young women participating in Prolific studies; see Letzter, 2021; again, also see <https://tinyurl.com/4unuvnb3> for information on reducing the likelihood of such issues). Acknowledging these considerations, our results may be helpful for guiding researchers interested in using Prolific for longitudinal research, and they suggest that sound data can be obtained quickly for study protocols involving relatively few longitudinal study assessments.

References

- Adamson, S. J., Kay-Lambkin, F. J., Baker, A. L., Lewin, T. J., Thornton, L., Kelly, B. J., & Sellman, J. D. (2010). An improved brief measure of cannabis use: The Cannabis Use Disorders Identification Test-Revised (CUDIT-R). *Drug and Alcohol Dependence, 110*, 137-143.
- Arditte, K. A., Cek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment, 28*, 684-691.
- Bauer, I. E., Meyer, T. D., Sanches, M., Spiker, D., Zunta-Soares, G., & Soares, J. C. (2017). Are self-rated and behavioural measures of impulsivity in bipolar disorder mainly related to comorbid substance use problems? *Cognitive Neuropsychiatry, 22*, 298-314.
- Bounoua, N., Church, L., & Sadeh, N. (2021). Alterations in reward and emotional processing differentiate among adults with a history of childhood maltreatment: Implications for substance use behaviors. *Emotion*. Advance online publication.
- Carpenter, R. W., Lane, S. P., Bruehl, S., & Trull, T. J. (2019). Concurrent and lagged associations of prescription opioid use with pain and negative affect in the daily lives of chronic pain patients. *Journal of Consulting and Clinical Psychology, 87*, 872-886.
- Chmielewski, M., Ruggero, C. J., Kotov, R., Liu, K., & Krueger, R. F. (2017). Comparing the dependability and associations with functioning of the *DSM-5* Section III trait model of personality pathology and the *DSM-5* Section II personality disorder model. *Personality Disorders: Theory, Research, and Treatment, 8*, 228-236.
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*, 1412-1427.

- Clark, L. A., & Watson, D. (1999). Temperament: A new paradigm for trait psychology. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality* (2nd ed., pp. 102–138). New York, NY: Guilford Press.
- Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, 32, 917-929.
- Everett, J. E. (1983). Factor comparability as a means of determining the number of factors and their rotation. *Multivariate Behavioral Research*, 18, 197–218.
- Hirschfeld, R. M. A., Williams, J. B. W., Spitzer, R. L., Calabrese, J. R., Flynn, L., Keck, P. E., Lewis, L., McElroy, S. L., Post, R. M., Rapoport, D. J., Russell, J. M., Sachs, G. S., & Zajecka, J. (2000). Development and validation of a screening instrument for bipolar spectrum disorder: The Mood Disorder Questionnaire. *American Journal of Psychiatry*, 157, 1873-1875.
- Jonason, P. K., & Webster, G. D. (2010). The Dirty Dozen: A concise measure of the Dark Triad. *Psychological Assessment*, 22, 420-432.
- Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. *Addiction*, 114, 609-619.
- Lane, S. P., & Hennes, E. P. (2019). Conducting sensitivity analyses to identify and buffer power vulnerabilities in studies examining substance use over time. *Addictive Behaviors*, 94, 117-123.

- Letzter, R. (2021). A teenager on TikTok disrupted thousands of scientific studies with a single video. *The Verge*. <https://www.theverge.com/2021/9/24/22688278/tiktok-science-study-survey-prolific>
- Mellis, A. M., & Bickel, W. K. (2020). Mechanical Turk data collection in addiction research: Utility, concerns and best practices. *Addiction, 115*, 1960-1968.
- Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*, 184-188.
- Roos, C. R., Kober, H., Trull, T. J., MacLean, R. R., & Mun. C. J. (2020). Intensive longitudinal methods for studying the role of self-regulation strategies in substance use behavior change. *Current Addiction Reports, 7*, 301-316.
- Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption—II. *Addiction, 88*, 791–804.
- Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment, 31*, 1428-1441.
- Shrout, P., & Lane, S.P. (2012). Psychometrics. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). Guilford Press.
- Stanton, K. (2021, October 1). Longitudinal prolific and substance use project. Retrieved from osf.io/x37ej
- Stanton, K., & Watson, D. (2016). Adult ADHD: Associations with personality and other psychopathology. *Journal of Psychopathology and Behavioral Assessment, 38*, 195–208.

- Stanton, K., & Watson, D. (2020). *The temporal stability and construct validity of measures assessing mania symptoms and risk: A comparative analysis*. Unpublished manuscript.
- Stanton, K., Khoo, S., McDonnell, C. G., Villalongo Andino, M., Sturgeon, T., & Aasen, L. (2021). An initial investigation of the joint classification of hypomania- and neurodevelopmental disorder-relevant dimensions within the Hierarchical Taxonomy of Psychopathology. *Assessment*. Advance online publication.
- Stasik-O'Brien, S. M., Brock, R. L., Chmielewski, M., Naragon-Gainey, K., Koffel, E., McDade-Montez, E., O'Hara, M. W., & Watson, D. (2019). Clinical utility of the Inventory of Depression and Anxiety Symptoms (IDAS). *Assessment*, 26, 944–960.
- Schick, M. R., Weiss, N. H., Contractor, A., Dixon-Gordon, K. L., Spillane, N. S. (2019). Depression and risky alcohol use: An examination of the role of difficulties regulating positive emotions in trauma-exposed individuals. *The American Journal of Drug and Alcohol Abuse*, 45, 323-332.
- Stone, A. A., Schneider, S., & Harter, J. K. (2012). Day-of-week mood patterns in the United States: On the existence of 'Blue Monday', 'Thank God it's Friday' and weekend effects. *The Journal of Positive Psychology*, 7, 306-314.
- Strickland, J. C., & Stoops, W. W. (2018). Feasibility, acceptability, and validity of crowdsourcing for collecting longitudinal alcohol use data. *Journal of the Experimental Analysis of Behavior*, 110, 136-153.
- Strickland, J. C., & Stoops, W. W. (2019). The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. *Experimental and Clinical Psychopharmacology*, 27, 1-18.

- Strickland, J. C., & Victor, G. A. (2020). Leveraging crowdsourcing methods to collect qualitative data in addiction science: Narratives of non-medical prescription opioid, heroin, and fentanyl use. *International Journal of Drug Policy*, 725, 102587.
- Tabish, S. A. (2020). COVID-19 pandemic: Emerging perspectives and future trends. *Journal of Public Health Research*, 9, 1786.
- Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, 38, 319-350.
- Watson, D., & Clark, L.A. (1999). *The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form*. Retrieved from Iowa Research Online: http://ir.uiowa.edu/psychology_pubs/11/
- Watson, D., Nus, E., & Wu, K. D. (2019). Development and validation of the Faceted Inventory of the Five-Factor Model. *Assessment*, 26, 17-44.
- Watson, D., O'Hara, M. W., Naragon-Gainey, K., Koffel, E., Chmielewski, M., Kotov, R., Stasik, S. M., & Ruggero, C. J. (2012). Development and validation of new anxiety and bipolar symptom scales for an Expanded Version of the IDAS (the IDAS-II). *Assessment*, 19, 399-420.
- Watts, A. L., Boness, C. L., Loeffelman, J. E., Steinley, D., & Sher, K. J. (2021). Does crude measurement contribute to observed unidimensionality of psychological constructs? An example with *DSM-5* alcohol use disorder. *Journal of Abnormal Psychology*, 130, 512-524.
- Weiss, N. H., Bold, K. W., Sullivan, T. P., Armeli, S., & Tennen, H. (2017). Testing bidirectional associations among emotion regulation strategies and substance use: A daily diary study. *Addiction*, 112, 695-704.

Wright, A. G. C., & Simms, L. J. (2014). On the structure of personality disorder traits: Conjoint analyses of the CAT-PD, PID-5, and NEO-PI-3 trait models. *Personality Disorders: Theory, Research, and Treatment*, 5, 43–54.

Table 1*Demographic Information for All Study Samples*

	Prolific Daily Diary Sample	Prolific Test-Retest Sample, Time 1	Prolific Test-Retest Sample, Time 2	Undergraduate Sample
Age (<i>M</i> & <i>SD</i>)	31.5 (10.1)	32.8 (11.7)	32.8 (11.8)	19.5 (1.7)
Gender				
Women	63.2%	63.9%	62.3%	74.3%
Men	28.7%	32.6%	34.2%	25.3%
Nonbinary	5.6%	3.5%	3.4%	0.2%
Transgender Women	0.9%	0.7%	0.8%	0.0%
Transgender Men	1.9%	1.2%	1.3%	0.2%
Sexual Orientation				
Heterosexual	60.8%	64.8%	63.7%	91.5%
Bisexual	26.2%	21.2%	22.5%	7.0%
Gay or Lesbian	9.9%	7.7%	7.7%	1.1%
Write-in Response	3.1%	6.3%	6.1%	0.4%
Racial Identity				
White or European American	82.9%	82.8%	83.6%	69.9%
Black or African American	4.0%	9.8%	9.3%	7.0%
Asian American	3.7%	7.2%	7.2%	14.6%
Ethnicity: Hispanic or Latino/a	12.5%	6.5%	6.4%	6.2%
Education				
High school diploma or less	13.1%	15.4%	14.9%	-
Some college or associate's	41.4%	39.9%	40.8%	-
Bachelor's degree	32.7%	29.8%	30.2%	-
Post-bachelor's education	12.8%	14.5%	13.5%	-
Employment				
Full Time	38.9%	38.7%	38.5%	-
Part Time	23.7%	20.5%	20.2%	-
Unemployed	37.4%	40.6%	41.1%	-

Note. $N = 321$ at Day 1 for the Daily Diary Sample; $Ns = 429$ and 377 at Times 1 and 2 in the test-retest sample; $N = 529$ for the undergraduate sample. Total percentages are not exactly 100% in some cases because participants could select multiple response options, write in other responses, or did not provide relevant information. Small percentages of participants also endorsed other racial identities in each sample, including being of Middle Eastern or North African Descent, Native American, Native Hawaiian or Pacific Islander, or a write-in response not included as a response option on our demographic questionnaire.

Table 2*Frequencies for Self-Reported Mental Health and Substance Use History Variables*

	Prolific Time 1	Prolific Time 2	Undergraduate
<i>Current Psychotherapy and Medication</i>			
Current psychotherapy	32.4%	31.0%	19.8%
Any psychiatric medication	52.9%	52.0%	18.3%
Antidepressants	42.0%	41.4%	13.0%
Antianxiety	19.8%	19.6%	5.5%
Stimulants	8.9%	7.4%	4.9%
Mood Stabilizers	5.4%	4.8%	0.8%
Antipsychotics	3.0%	3.2%	1.1%
<i>Other Psychosocial Information</i>			
Ever sought substance use treatment	10.3%	8.8%	2.1%
Ever hospitalized for mental health issues	29.8%	27.9%	5.5%
Ever been arrested	19.6%	18.8%	0.9%
<i>Self-Identified Lifetime Diagnoses</i>			
Alcohol or Other Substance Use Disorder	9.3%	8.2%	1.5%
Any Depressive Disorder	71.8%	71.6%	17.6%
Generalized Anxiety Disorder	60.1%	58.1%	20.0%
Social Anxiety Disorder	30.3%	31.0%	6.0%
Attention-Deficit/Hyperactivity Disorder	19.8%	17.8%	7.8%
Panic Disorder	17.2%	17.0%	4.2%
Bipolar Spectrum Disorder	11.0%	10.1%	1.3%
Obsessive Compulsive Disorder	9.1%	9.3%	4.9%
Any Eating Disorder	8.4%	9.0%	4.0%
Autism Spectrum Disorder	6.5%	7.2%	0.2%
Any Personality Disorder	5.8%	5.3%	0.8%
Schizophrenia or Psychotic Disorder	2.1%	1.6%	0.2%

Note. $N = 429$ and 377 , respectively in the Prolific sample across timepoints; $N = 529$ for the undergraduate sample. Sample sizes were slightly smaller by one to several cases for some variables due to missing data. The substance use treatment seeking item asked participants about seeking treatment for alcohol use or any other substance.

Table 3
Descriptive Statistics, Frequency Information, and Mean Differences for Measures Across Samples and Timepoints

	<u>Prolific Time 1</u>	<u>Prolific Time 2</u>	<u>T1-T2 <i>d</i></u>	<u>Undergraduate</u>	<u>T1 <i>d</i></u>
	<i>M (SD) and ω</i>	<i>M (SD) and ω</i>		<i>M (SD) and ω</i>	
<i>Substance Use</i>					
AUDIT	5.2 (6.7); .92	4.7 (6.3); .91	.08	5.5 (5.4); .85	-.05
CUDIT-R	3.3 (5.7); .90	2.9 (5.2); .89	.07	2.6 (4.9); .90	.13
Drug Use Survey	7.9 (9.6); .89	7.5 (9.5); .90	.04	2.3 (3.5); .77	.78
<i>AUDIT Classification (%)</i>					
Low	76.7%	79.6%	-	67.6%	-
Hazardous	14.9%	12.7%	-	26.7%	-
Very Hazardous	2.8%	4.2%	-	3.8%	-
Severe	5.6%	3.5%	-	1.9%	-
<i>CUDIT-R Classification (%)</i>					
Low	82.1%	83.8%	-	87.5%	-
Problems	7.7%	8.0%	-	4.7%	-
Probable Use Disorder	10.3%	8.2%	-	7.8%	-

(table continues)

Table 3 (cont.)

	<u>Prolific Time 1</u>	<u>Prolific Time 2</u>	T1-T2 <i>d</i>	<u>Undergraduate</u>	T1 <i>d</i>
	<i>M (SD) and ω</i>	<i>M (SD) and ω</i>		<i>M (SD) and ω</i>	
<i>Hypomania and Internalizing</i>					
IDAS-II Dysphoria	28.3 (9.1); .90	27.6 (9.3); .90	.08	26.1 (9.6); .91	.24
Mood Disorder Questionnaire	6.8 (4.0); .91	6.2 (4.2); .90	.15	7.3 (3.6); .85	-.13
<i>Positive Symptom Screening</i>					
IDAS-II Any Internalizing	50.8%	45.4%	-	39.5%	-
Mood Disorder Questionnaire	19.6%	19.6%	-	10.0%	-
<i>Personality Pathology</i>					
Disinhibition	19.9 (10.5); .91	19.3 (10.2); .91	.06	15.5 (8.4); .88	.46
Antagonism	9.3 (8.8); .91	9.3 (8.8); .92	.00	8.1 (7.0); .87	.15
Internalizing	30.0 (11.1); .91	29.5 (11.0); .90	.04	24.2 (11.2); .91	.52
Psychoticism	7.6 (8.9); .90	6.6 (8.7); .91	.11	8.5 (7.9); .87	-.11
Detachment/Social Withdrawal	13.0 (5.8); .86	13.2 (5.7); .85	-.03	8.6 (4.7); .79	.83

Note. $N = 429$ and 377 , respectively in the Prolific sample across timepoints; $N = 529$ for the undergraduate sample. All effect size differences shown are Cohen's d values. The set of Cohen's d values in the final column are those representing differences between the Prolific Sample at Time 1 and the Undergraduate sample. Cohen's d values $\geq .20$ representing at least a small effect size are **bolded**. All personality pathology measures are based on scales from the Comprehensive Assessment of Traits Relevant to Personality Disorder. AUDIT = Alcohol Use Disorders Identification Test; CUDIT-R = Cannabis Use Disorders Identification Test-Revised; IDAS-II = Expanded Version of the Inventory of Depression and Anxiety Symptoms. Mood Disorder Questionnaire mean scores represent the sum of the 13 item scores included in this questionnaire.

Table 4*Test-Retest Stability Coefficients for Study Scales Over a 2-Week Interval*

Scale	Retest Correlation	95% Confidence Interval
<i>Substance Use</i>		
AUDIT	.93	.92 to .94
CUDIT-R	.90	.88 to .92
DUS	.95	.94 to .96
<i>Other Symptoms</i>		
Mood Disorder Questionnaire	.80	.76 to .83
IDAS-II Dysphoria	.87	.84 to .89
<i>Personality Pathology</i>		
Disinhibition	.91	.89 to .93
Antagonism	.87	.84 to .89
Internalizing	.90	.88 to .92
Psychoticism	.86	.83 to .88
Detachment/Social Withdrawal	.87	.84 to .89

$N = 377$. Values shown are Pearson correlations, and all correlations were significant at a $p < .0001$ level. AUDIT = Alcohol Use Disorders Identification Test; CUDIT-R = Cannabis Use Disorders Identification Test-Revised; IDAS-II = Expanded Version of the Inventory of Depression and Anxiety Symptoms. Mood Disorder Questionnaire, AUDIT, CUDIT-R, and IDAS-II scores used in these analyses were the sum of item scores from each of these respective questionnaires rather than positive screening percentages.

Figure 1

Graphical representation of alcohol use (percentage of days with at least one drink, average number of drinks on days with drinking), negative and positive affect, fatigue, and number of negative interpersonal experiences over the 5-day daily diary protocol. All participants started on Wednesday and finished on Sunday.

Online Appendix S1

Daily Diary Study Overview and Rationale for Data Exclusions

The data presented in the demonstrative analyses for this first study were drawn from a broader study focused on understanding associations for (a) daily alcohol use and use of other substances with (b) daily ratings mood, interpersonal experiences (e.g., experiences of conflict), and a range of mental health symptoms. Our baseline battery administered at Day 1 included assessment of a range of demographic variables and the baseline measures of substance use described in the main text (e.g., alcohol use over the past 6 months), as well as a very large number of personality and psychopathology variables not incorporated into our primary study analyses (e.g., over 70 items assessing hypomania history and/or risk; over 130 items assessing personality pathology; multiple measures of internalizing symptoms). This extensive initial battery of measures was administered to enable us to examine associations for these variables with daily ratings over the 5-day study period. Additionally, over 100 items were administered daily. This included assessment of daily exercise and other activities (e.g., when participants first had social contact, exercise and eating patterns, experiences of physical pain) in addition to the daily ratings reviewed in the main text focusing on mood, experiences of interpersonal conflict, and so on.

It would have been very challenging to integrate all scores from these various measures in a manuscript focused solely on this dataset, let alone in the context of the current study described in the main text that also includes data from two other samples. Additionally, some measures were outside of the scope of the current study (e.g., ratings of how many times participants left their homes on a given day), which focused on illustrating how Prolific can be used to collect data to address contemporary topics in substance use and mental health research when using

repeated measures designs (e.g., improving understanding of how daily experiences of interpersonal conflict are related to the occurrence and amount of alcohol use on a given day).

Prolific Test-Retest Study Overview and Rationale for Data Exclusions

The test-retest data from the second Prolific sample and the undergraduate sample described in the main text were collected for a broader study focused on informing dimensional symptom classification and assessment from the perspective of the Hierarchical Taxonomy of Psychopathology. Data from a range of substance use, personality pathology, hypomania, and internalizing symptom measures already are presented in the main text, as are extensive data on participants' demographic backgrounds and mental health histories. However, we did not include other specific demographic data in our analyses here (e.g., disruptions caused by COVID; more specific details of employment and legal histories) due to length constraints. Data from other measures that we could not incorporate due to space limitations include ratings of traits relevant to neurodevelopmental disorders (e.g., inattentiveness, hyperactivity, social communication differences), including retrospective ratings of these and other traits in childhood (e.g., “as a child did you have academic difficulties”; “were you prone to being irritable”).

Study data and associated materials for both samples are available on the Open Science Platform as described in the main text. Please contact the authors for additional information about these studies and datasets.

Online Supplemental Table S1*Factor Loadings for Substance Use Items Across Samples and Timepoints*

	<u>Prolific Time 1</u>	<u>Prolific Time 2</u>	<u>Undergraduate</u>
<u>AUDIT</u>			
3. Frequency of cannot stop drinking	.95	.95	.81
4. Failed responsibilities due to drinking	.92	.91	.77
6. Feelings of guilt after drinking	.91	.90	.73
5. Needed a drink to get going in morning	.89	.88	.74
2. Frequency of 6+ drinks	.89	.90	.86
7. Blackouts due to drinking	.88	.89	.86
8. Number of drinks typically when drinking	.83	.85	.82
10. Others concerned about drinking	.75	.73	.52
1. Frequency of having an alcoholic drink	.73	.72	.79
9. Self or others injured due to drinking.	.55	.60	.61
<u>CUDIT-R</u>			
6. Great deal of time using/getting cannabis	.93	.90	.91
1. Frequency of cannabis use	.93	.91	.98
7. Memory or concentration difficulties	.90	.93	.94
4. Trouble stopping use once starting	.90	.88	.90
5. Failed responsibilities due to use	.88	.89	.91
2. Hours high on a typical day of use	.87	.88	.87
3. Ever thought of reducing use	.86	.91	.82
8. Used cannabis in dangerous situations	.84	.81	.90
<u>Drug Use Survey</u>			
2. Cocaine	.89	.89	.94
4. Opioids or painkillers	.87	.86	.57
7. Ecstasy	.85	.87	.92
3. Amphetamines	.84	.91	.77
6. Psychedelics	.84	.86	.84
5. Tranquilizers	.80	.83	.70
1. Cannabis	.76	.74	.84
8. Stimulants such as Ritalin	.76	.78	.89

$N = 429$ and 377 , respectively, at Time 1 and Time 2 in the Prolific sample and 529 in the undergraduate sample for all analyses. AUDIT = Alcohol Use Disorders Identification Test; CUDIT-R = Cannabis Use Disorders Identification Test-Revised.