Validity of a Two-Item Screen for Early Psychosis

Peter L. Phalen , Pamela Rakhshan Rouhakhtar , Zachary B. Millman , Elizabeth Thompson , Jordan DeVylder , Vijay Mittal , Evan Carter , Gloria Reeves , Jason Schiffman

PII: S0165-1781(18)31480-X
DOI: https://doi.org/10.1016/j.psychres.2018.11.002
Reference: PSY 11858

To appear in: *Psychiatry Research*

Received date: 10 August 2018
Revised date: 2 November 2018
Accepted date: 2 November 2018

Please cite this article as: Peter L. Phalen , Pamela Rakhshan Rouhakhtar , Zachary B. Millman , Elizabeth Thompson , Jordan DeVylder , Vijay Mittal , Evan Carter , Gloria Reeves , Jason Schiffman , Validity of a Two-Item Screen for Early Psychosis, *Psychiatry Research* (2018), doi: https://doi.org/10.1016/j.psychres.2018.11.002

HIGHLIGHTS

- We introduce a very brief (2-item) and simple (yes-no) screen for early psychosis
- The measure had sensitivity/specificity comparable to other widely used brief screens
- May be especially useful in settings with low buy-in (schools, primary care, &c.)
- We use Bayesian Item-Response Theory models and provide model code
- Using quantified uncertainty, we also identify/validate alternative 2-item screens

Validity of a Two-Item Screen for Early Psychosis

Peter L. Phalen[a,b], Pamela Rakhshan Rouhakhtar[c], Zachary B. Millman[c], Elizabeth Thompson[d], Jordan DeVylder[e], Vijay Mittal[f], Evan Carter[g], Gloria Reeves[a], and Jason Schiffman[c]*

[a] Department of Psychiatry,
University of Maryland School of Medicine
Baltimore, Maryland, USA

[b] Mental Illness Research, Education and Clinical Center (MIRECC),
Baltimore VA Medical Center
Baltimore, Maryland, USA

[c] Department of Psychology,
University of Maryland, Baltimore County
Baltimore, Maryland, USA

[d] Department of Psychiatry and Human Behavior
Brown University, Alpert Medical School
Providence, Rhode Island, USA

[e] Graduate School of Social Service
Fordham University
New York, New York, USA

[f] Department of Psychology
Northwestern University
Evanston, Illinois, USA

[g] Human Research and Engineering Directorate (HRED)
United States Army Research Laboratory
Adelphi, Maryland, USA

*Corresponding Author:
Jason Schiffman
Department of Psychology
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
schiffma@umbc.edu

Abstract

Well-validated screening tools have been developed to identify people at high risk for psychosis, but these are rarely used outside of specialty clinics or research settings. The development of extremely brief and simple screening tools could increase dissemination, especially in settings with low buy-in such as those with low base rates of psychosis and/or time constraints. We sought to identify such a brief measure by modeling participant responses to three psychosis screening questionnaires (Prime Screen; Prodromal Questionnaire-Brief; Youth Psychosis At Risk Questionnaire) in a sample of 139 help-seeking individuals and 335 college students (age range: 12-25). Two screening questions with especially strong information characteristics were identified: "Do you see things that others can't or don't see?" and "Have you ever felt that someone was playing with your mind?" (Alternative two-item screens with similarly strong properties were also identified and validated using uncertainty quantified through Bayesian modeling.) The resulting measure was validated against clinician ratings of psychosis. The screen performed with a sensitivity of 53% and specificity 98% for clinically significant hallucinations or delusions, and sensitivity of 32% and specificity 99% for identifying people in an early phase of psychosis (clinical high risk or first episode psychosis).

1. Introduction

Prognosis for individuals at risk of developing a psychotic disorder may be improved through early intervention. Screening tools have been developed to identify individuals with emerging psychosis or clinical high risk symptoms as a strategy to reduce the duration of untreated psychosis (Schiffman, 2018). Validated screeners designed for this purpose include the Prodromal Questionnaire-Brief (PQ-B), the Prime Screen, and the Youth Psychosis At Risk Questionnaire (YPARQ). These screeners align well with clinician ratings of psychosis (AUCs

typically measured at around 0.8; Kline et al., 2012) and may be especially useful in clinical settings with high false positive rates (Fusar-Poli et al., 2016).

Despite promising psychometric properties and their relative brevity and ease of use compared to clinical interviews, widespread implementation of these validated measures has not taken hold. Given the relatively low base-rates of psychosis, community workers may be reluctant to administer psychosis-specific screening measures as they may doubt the relevance. Psychosis is often just one of many conditions under consideration, in which case screening tools for a single disorder can be either too specific or too long relative to those necessary to screen for other concerns. Additionally, the busy nature of many settings (e.g., schools, primary care, ER) elevates efficiency to primary importance, making certain item styles (e.g., 6 point Likert scales) that are common in existing psychosis-specific screeners, and afford advantages in some contexts, less attractive in these settings. Therefore, despite the advantages of existing screening measures, it may be that the costs simply outweigh the perceived benefits (Savill et al., 2018).

In contrast, rapid and simple screening measures may preserve some important benefits of existing measures while limiting the costs. Very brief (e.g., two items) and simple (e.g., yes/no) measures could even be memorized and delivered verbally without the need for paper or electronic materials, thereby drastically reducing burden for both clinicians and clients. Without displacing existing measures, such a brief tool could be implemented in settings with very low buy-in that would not otherwise administer any formal screening tools at all. To our knowledge there has been one previous attempt to screen for psychosis using a very brief (single item) screening tool with simple scoring (Kelleher et al., 2011). This screening tool, however, was not validated against a measure of clinically significant psychotic symptoms and was developed exclusively in a non-clinical sample.

In the present study, we analyzed patterns of yes-no responses to questions from three existing psychosis screening tools (the Prime Screen, PQ-B, and Y-PARQ) in a combined sample of clinical and college-based adolescents and young adults. The clinical sample was assessed using a clinician-rated diagnostic interview for early psychosis and rated for the presence of hallucinations or delusions that affect daily functioning. Using analytical methods derived from item response theory (IRT), we aimed to identify a two-item screener with a strong ability to distinguish between respondents at a level of psychosis severity consistent with the presence of delusions or hallucinations that affect daily functioning. We then assessed the validity of this two-item screen by checking the measure's ability to discriminate between people with versus without clinically significant delusions/hallucinations, as well as people with versus without a diagnosis within the early psychosis spectrum (clinical high risk or first episode psychosis). We provide a rationale for our preferred score threshold and its associated specificity/sensitivity, while also reporting sensitivity and specificity for all possible cut points to allow clinicians or researchers to select their preferred threshold if their reasoning or purposes diverge from those considered here.

Our use of a combined clinical and college sample improves estimation of the item response models which do not make assumptions about clinical outcome, and is reasonable given that the population for which such a small screening tool is most likely to be implemented would just as likely resemble the college sample (low risk, active psychosis unlikely) as the clinical sample. There are reliable demographic differences, however, between college and clinical samples which may confound results. Therefore, we also test for differential item functioning by group, and present receiver operating characteristics for the combined clinical and college sample as well as results when the sample is restricted to the clinical group only.

2. Materials and Method

2.1 Participants

This study was conducted through the Youth FIRST research program/Strive for Wellness clinic affiliated with the University of Maryland, Baltimore County (UMBC) and the University of Maryland School of Medicine. The full sample (N = 474) was comprised of help-seeking (i.e., clinical) participants (N = 139) and college participants (N = 335). Clinical participants were help-seeking adolescents and young adults between the ages of 12-25, recruited within the community and actively receiving mental health services. College participants were undergraduate students recruited between November 2010 and May 2014 from introductory Psychology courses at the University of Maryland, Baltimore County (UMBC). Inclusion criterion for the undergraduate study was age greater than 18 at the time of participation. There were no exclusion criteria.

[Insert Table 1 about here]

Thirty-six college and 29 clinical participants were missing at least one item (46 people missing one item, 14 missing two or three items, and 5 missing between four and six items). Some of these items were removed after preliminary factor analyses revealed poor loading (described below). No participant was missing more than 3 of the items that were ultimately analyzed in the IRT model, therefore, no participants were excluded from analyses due to missing data.

2.2 Measures

2.2.1 Structured Interview for Psychosis Risk Symptoms (SIPS). The SIPS is a semi-structured, clinician-administered interview that measures the presence and severity of symptoms associated with early psychosis (Miller et al., 2003). It is among the most widely used measures

of clinical high risk for psychosis, representing the gold standard for assessment of psychosis-risk (Fusar-Poli et al., 2016). The presence or absence of three risk syndromes (attenuated psychosis syndrome, brief intermittent psychosis syndrome, and genetic risk syndrome), as well as full threshold psychosis is determined from the SIPS interview protocol. Participants were designated at "clinical high-risk" (CHR) if they were positively diagnosed with a risk disorder or schizotypal personality disorder. Clinician-rated scores of 5 or greater on the SIPS hallucinations or delusions scales were categorized as experiencing active hallucinations or delusions. The anchor for a SIPS rating of 5 for hallucinations is, "Hallucinations experienced as external to self though skepticism can be induced by others. Mesmerizing, distressing. Affects daily functioning." The anchor for a SIPS rating of 5 for delusions is "Experiences familiar, anticipated. Doubt can be induced by contrary evidence and others' opinions. Distressingly real. Affects daily functioning."

SIPS interviews were only completed for clinical participants. For analyses involving both college and clinical participants, it was assumed that participants in the college sample did not have a psychotic disorder and were not considered actively psychotic. We believe this assumption is statistically reasonable given that available evidence suggests an approximate 2.5% incidence rate for psychotic experiences in the general population (Linscott and van Os, 2013), with prevalence of clinically significant experiences, and prevalence in college samples, still lower. In the present sample, a 2.5% rate of psychotic experiences would correspond to 8 college students, representing a likely upper bound on how many people may be misclassified.

2.2.2 Prime Screen Revised. The Prime Screen Revised was developed by the authors of the SIPS (Miller et al., 2004). The Prime Screen consists of 12 Likert-type items that describe attenuated positive psychosis symptoms. Participants are asked to choose from 7 response options, from "definitely disagree" (0) to "definitely agree" (6). For the current study, responses

of "slightly agree" was considered a minimally affirmative response and therefore coded as a yes answer, whereas responses of "not sure" or lower were recoded as no answers.

2.2.3 Prodromal Questionnaire-Brief. The Prodromal Questionnaire-Brief (PQ-B) is a 21 item self-report scale, developed from the original 92 item Prodromal Questionnaire (Lowey et al., 2005), measuring presence and severity of attenuated positive psychosis symptoms (Lowey et al., 2011). Participants indicate presence or absence of attenuated psychosis symptoms through "Yes/No" responses. In the event of a positive endorsement, participants are asked to rate the extent to which the endorsed symptoms causes concern, fright, or impairment, ranging from a score of one ("strongly disagree") to a five ("strongly agree"). The distress score was not used in this study given that our aims were to identify binary items that are most informative regarding psychosis status.

2.2.3 Youth Psychosis At-Risk Questionnaire-Brief. The Youth Psychosis at Risk Questionnaire-Brief (YPARQ-B) is a 28-item self-report questionnaire measuring presence of attenuated positive symptoms that are associated with prodromal schizophrenia (Ord et al., 2004). Participants indicate the presence or absence of attenuated psychosis symptoms by responding "yes," "no," or "unsure." For the present study, responses of unsure were coded as "no" answers, which is in line with the scale creators' recommendations (Ord et al., 2004).

2.3 Procedure

Procedures were approved by Institutional Review Boards at UMBC, the University of Maryland School of Medicine, and the Maryland Department of Health IRB. Prior to completion of study procedures, all participants (and guardians, in the case of underage participants) received an overview of the study and consented to their involvement. Across both studies, participants completed a self-report demographics form and the three psychosis screening tools described above.

Participants in the clinical sample—but not the college sample—were administered the SIPS by study staff. All SIPS administrators underwent a two-day training workshop led by one of the original authors of the SIPS (Barbara C. Walsh) or were trained by trainees of that author. Additionally, study staff co-rated audio-recorded and live interviews, and were observed administrating the SIPS by experienced raters. Raters were approved to independently administer SIPS interviews only when reliability coefficients were 0.80 or greater. All cases were reviewed weekly via case conferences with experts in the assessment of early psychosis. The within-lab ICC for SIPS positive symptoms was 0.82, and diagnostic agreement was $\kappa = 1$ (Millman et al., 2018). For the current study, SIPS administrators were blind to screening questionnaire responses prior to administration of the SIPS clinical interview.

2.4 Analysis.

Question preparation. Because the brief screener was intended to be as simple as possible (potentially even delivered verbally), and to increase comparability across screening measures, a yes-no format for responses was desirable. Responses to the Prime Screen were therefore recoded dichotomously to mimic a yes-no answer by coding any minimally affirmative responses ("slightly agree" or higher) to 1 and all responses of "not sure" or lower to 0. Responses to the PQ-B and the Y-PARQ are already recorded in a yes-no format and therefore did not require transformation.[1]

Item Response Theory (IRT) models and item selection. IRT models can be used to quantify the precision with which test items identify people with varying levels of a latent construct measured by a test (in this case, the latent construct is psychosis severity). IRT models

[1] We compared the effects of recoding the Prime Screen responses of 'agree' and above versus 'slightly agree' and above as 'yes' by fitting IRT models using both cut points. We found that coding only 'agree' responses and above as 'yes' yielded low affirmative response rates on the Prime as compared to the same participants' rates of affirmative responding on the other screens, despite manifest similarities in the contents/phrasing of the items between measures. In contrast, coding 'slightly agree' and above as a 'yes' on the Prime corresponded well to response rates on the YPARQ and the PQ-B.

generally assume that all items capture a common factor. Therefore, a principal component analysis was performed before fitting the models. Items with high loadings on the primary identified factor (0.4 or greater) were selected for further analysis (Hinkin 1998).

The IRT model in this study was specified and fit using Stan, a flexible programming language that enables a fully Bayesian approach (Stan Development Team 2017). The model is a hierarchical 2-parameter item-response with psychosis severity as the latent construct captured by the items. We concurrently estimated average psychosis severity for people with versus without clinician-rated hallucinations or delusions by varying a severity intercept by group. Weakly informative priors were placed on all parameters to reduce risk of overfitting (cf. Furr, 2016). Differential item functioning by group (clinical versus college), as well as interactions between group and key demographic factors (age, race, and gender), were assessed by varying item parameters by group and demographic for each item (see Furr et al., 2016). However, no item performed significantly differently in the clinical versus the college sample. Therefore, parameters for differential item functioning were dropped from the final model. The full code for the final model is provided in an Online Supplement.

Items were selected based on their "information" (see Mead and Meade, 2010), a measure of precision equal to 1/SE(severity). Items with higher item information at a given latent severity level are better able to identify people at that level of severity. We selected the two items with the highest item information at the median level of latent psychosis severity estimated for people with active hallucinations or delusions, while also identifying alternative items that have a not insubstantial (at least 25%) probability of being in the "true" top two at the population level.

Criterion and construct validity. ROC curves were constructed for the resulting two-item measure, with CHR or Psychotic Disorder and active hallucinations or delusions as outcome measures based on the SIPS administered in the clinical sample. We constructed separate ROC

curves for the clinical group and for the combined clinical/non-clinical sample in order to better understand how the statistic was affected in samples with high base rate. As noted above, SIPS interviews were only completed for clinical participants and it was assumed that members of the college sample would not reach clinical criteria for psychosis. Therefore, measured sensitivity was necessarily identical in the clinical and combined clinical/college samples, but there was the potential for substantial variation in specificity with the negative college cases.

As described above, we believe the assumption that college students are negative cases is statistically reasonable given that available evidence suggests an approximate 2.5% incidence rate for psychotic experiences in the general population, with prevalence of clinically significant experiences necessarily lower, and prevalence in college populations lower still (this corresponds to a likely upper bound of 8 misclassifications out of 335 college participants; Linscott & van Os, 2013). The lack of a clinical psychosis interview in this subsample, however, is a potential concern. Therefore, to further test the validity of the two-item scale in the college sample, we estimate correlations between the identified two-item screen and the following self-report measures: the Schizotypal Personality Questionnaire (SQP), the Beck Depression Inventory-II (BDI-II), the Beck Anxiety Inventory-II (BAI-II), as well as total scores on the three psychosis screeners described above (Prime, PQ-B, and YPARQ-B).

3. Results

The SIPS diagnostic interview performed on the 139 clinical participants identified 71 positive cases: 47 Clinical High Risk and 24 meeting criteria for Psychotic Disorder. The high rate of psychosis compared to the general population is partially due to the higher rates of psychosis measured in clinical samples and partially due to the fact that some community referrals were made due to the nature of the study and suspicion of psychotic symptoms. Eighteen participants were classified as experiencing active delusions (SIPS P1 $\geq$ 5) and 34

participants were identified as experiencing active hallucinations (SIPS P4 ≥ 5), with a total of 36 participants who were experiencing either active hallucinations or active delusions. Participant characteristics are available in Table 1.

3.1 IRT model and item selection

Principal components analysis identified one predominant factor for the 61 screening items. Any item with loading less than 0.4 onto this primary factor was dropped, leaving 49 items. These 49 items were then analyzed using a hierarchical 2-parameter item-response model (see Online Supplement for final model). There were no divergent transitions, and model diagnostics were within normal limits. Differential item functioning was assessed to identify any item that showed differing properties in the college versus the clinical group. No item showed significant variation by group, though this may have been attributable to the relatively small sample size of the clinical group which led to wide credible intervals for the subsample. Item 22 from the Y-PARQ ('Do you see things that others can't or don't see', see below) showed some evidence of better discrimination when applied to the clinical group than when applied to the college group (91% probability; i.e., 80% uncertainty intervals excluded zero). Eighty percent credible intervals did not exclude zero for any other item on either parameter. There was no significant evidence of interactions between group and gender, between group and race, or between group and age, on item functioning. Although differences exist between the clinical and non-clinical samples, the lack of evidence of demographic differences on item functioning do not suggest that combining samples for the purposes of this manuscript was inappropriate.

Figure 1 gives the distribution of estimated latent psychosis severity for each participant in the sample. As would be expected, there is a skew such that fewer participants have higher estimated severity levels, and estimated severity scores are higher on average for people with active hallucinations or delusions (>99.99% probability).

Figure 2 shows information curves for all 49 items. Information curves give the precision with which an item is able to identify people with varying degrees of psychosis severity. The two items with highest information at the estimated median severity (curves intersecting at highest point on the dotted line) for people with active hallucinations or delusions were selected. These were items 22 and 28 from the Y-PARQ. Respectively:

Do you see things that others can't or don't see?

Have you ever felt that someone was playing with your mind?

3.2 Criterion and construct validity.

For each participant in the sample, scores on the two-item screener were calculated as a simple addition of yes-no responses (e.g., two "yes" answers yield a score of 2 on the screener). Table 2a gives receiver operating characteristics for the two-item screen in the full clinical and college sample (*N*=471; note that three participants had missing answers for one of the items and were subsequently dropped from these analyses). An AUC of 0.77 was achieved for predicting CHR status or psychotic disorder and an AUC of 0.86 for predicting active hallucinations or delusions in the combined clinical/college sample. For comparison, in the same combined sample, the PQ-B, Prime, and Y-PARQ measures with full scoring protocols respectively achieved AUCs of 0.81, 0.86*, and 0.92* for CHR or psychotic disorder, and 0.89, 0.91, and 0.95* for hallucinations or delusions on the same sample.[2] (Recall that for all these analyses, college students were coded as non-psychotic.) A cut point of 2 rules out 98% of people without active hallucinations or delusions and catches 53% in the combined sample. The same cut point rules out 99% of people who are not at CHR or who meet criteria for psychotic disorder and catches 32% of those who are at High Risk or who meet criteria for psychotic disorder.

Table 2b gives receiver operating characteristics for the two-item screen restricted to the clinical sample (*N*=139). An AUC of 0.67 was achieved for predicting CHR status or psychotic disorder and an AUC of 0.78 for predicting active hallucinations or delusions. For comparison, in the same sample, the PQ-B, Prime, and Y-PARQ measures with full scoring protocols respectively achieved AUCs of 0.73, 0.73, and 0.79, for CHR or psychotic disorder, and 0.81*, 0.79, and 0.82* for hallucinations or delusions. A cut point of 2 yes answers on the two-item measure rules out 95% of people at low risk of psychotic disorder and catches 32% of those who

[2]Statistically significant ($p<0.05$) improvements over the performance of the 2-item measure are marked with an asterisk (*).

are at CHR or meet criteria for a psychotic disorder. The same cut point rules out 93% of people without active hallucinations or delusions and catches 53%.

As we did not deliver a clinical interview for the college sample, there were potential concerns about the construct validity of the two-item screen when restricted to that sample. To test the construct validity of the two-item screen in the college sample, however, we tested correlations between the two-item screen and the SPQ, BDI-II, BAI-II, and total scores on the three screeners (see Table 3). We found that the two-item screen correlated with each measure in the expected direction at $p<.0001$, with correlations between the brief two-item screen and the SPQ, BDI-II, and BAI-II frequently stronger than those exhibited by the full screening measures total scores.

3.3 Accounting for uncertainty.

To account for uncertainty and illustrate the value of the current method over the value of any two specific items, we identified any item that did not make it into the top two in our sample but had a reasonably high probability of being in the "true" top two at the population level. Specifically, we identified items that had a 25% probability of being in the "true" top two at the population level (no item had greater than 35% chances). The following three items met this criterion:

Y-PARQ 12: Do you ever hear the voice of someone talking that other people cannot hear?

Y-PARQ 20: Do familiar surroundings sometimes seem unreal to you?

PQ-B 20: Have you seen things that other people can't see or don't seem to see?

Each of these items had at least a 25% chance of being better than Y-PARQ item 28, while none of them credibly performed better than Y-PARQ item 22 at the population level. In the Appendix we give receiver operating characteristics for four credible two-item screeners

composed of the best performing item (Y-PARQ 22) plus each of the four next best performing items (i.e., 22 and 28; 22 and 12; 22 and 19; 22 and 20).

Practitioners may have preferences between the various potential two-item screens. Of note, arguably the most face valid two-item psychosis screener, composed of the Y-PARQ 12 and 22 ("do you ever hear the voice of someone talking that other people cannot hear?" and "do you see things that others can't or don't see?"), performs reasonably well. For example, using a cut point of two "yes" answers in the full clinical/non-clinical sample, this screen achieves 67% sensitivity and 96% specificity for active hallucinations or delusions. Additionally, not all two-item screeners perform equally well. For example, the screener composed of the Y-PARQ 22 ("Do you see things that others can't or don't see?") and the PQ-B 20 ("Have you seen things that other people can't see or don't seem to see?") has poor specificity in the clinical sample, perhaps due to the substantial semantic overlap between the two items, as well as the relatively high baseline rate of endorsement for the PQ-B 20 (which queries for lifetime as opposed to current incidence of visual hallucinations). Thus, while each of the featured two-item screening tools perform fairly well, attention should be paid to the particular receiver operating characteristics of each potential measure when deciding on implementation.

4. Discussion

Very brief and simply scored psychosis screening tools may facilitate screening in populations with low base rates of psychosis and environments with strong time constraints (e.g., primary care settings), and in scenarios where screening tools must be delivered without access to electronic health screeners (e.g., school-based clinics, low resourced settings). This study provides preliminary evidence for the validity of extremely brief 2-item yes-no questions as a screen for early psychosis. Validity was supported by the favorable operating characteristics of the two-item measure when predicting psychosis as determined by clinical interview performed

with help-seeking youth. The measure performed well against a combined clinical/non-clinical sample and when used with clinical participants only, although specificity decreased by about 4 percentage points in the purely clinical sample. The sensitivity and specificity characteristics described in Table 2 are comparable to other common very brief mental health symptom screening tools, such as the six-item Likert style K-6 screen for "serious mental illness" (sensitivity 36%; specificity 96%; with serious mental illness defined as any 12-month DSM-IV disorder other than a substance use disorder and a GAF score of less than 60; Kessler et al., 2003) or the two-item Likert style PHQ-2 screen for depression (sensitivity 83%, specificity 90%; Kroenke et al 2003).

At a cut point of two "yes" answers, the current two-item screener has a sensitivity of 53% for hallucinations or delusions that affect daily functioning and a specificity of 98% in the combined clinical/non-clinical sample.

Clinicians and/or researchers could select an appropriate cut point (either 1 or 2) depending upon desired sensitivity and specificity. (We report sensitivity and specificity associated with each cut point in Table 2.) When screening in the general population, there are compelling reasons to consider a cut point of two 'yes' answers. In our mixed clinical/nonclinical sample, estimated specificity for two 'yes' answer is 98%, versus 90% for a cut point of one. This 8-point difference in specificity between thresholds may seem negligible, however, given base rates of approximately 2.5% for psychotic symptoms in the overall population (Linscott and van Os, 2013), adopting a cut point of 2 rather than 1 would prevent almost 8 times the number of false positives than would be gained in true positives. For example, assuming a 2.5% rate of psychotic symptoms and the characteristics described in Table 2a, if 10,000 people were screened, adopting a cut point of 1 would yield 55 false negatives and 975 false positives, whereas a cut point of 2 would yield 118 false negatives and only 195 false

positives. When screening in a clinical sample with higher base rates of psychosis it may be more beneficial to sacrifice specificity for increased rates of detection. Still, a cut point of two may be preferable given that a threshold of only one 'yes' answer has very low specificity of <70% in the clinical cohort (versus >90% for a cut point of two).

Despite some disadvantages, longer screening tools would still be preferable in many situations, either alone or as a follow-up to a positive on a shorter screening protocol. When psychosis is suspected, longer measures can give more detailed information and detect an array of possible psychotic features. When the intent is to determine a diagnosis, no screening tool can replace a full clinical interview. Still, in many settings the goal is not to establish a final diagnosis or to monitor the degree of psychosis severity, but rather to take a "first step" and determine the likely need for any further assessment for the disorder in question. A brief screen may be optimal for settings of this kind, for example, when psychosis is not likely (intakes at non-specialty clinics), when many potential conditions have to be screened quickly due to time limitations, or when a provider prefers to complete screening verbally as part of a clinical interview.

Several considerations should be borne in mind when interpreting our findings. First, the sample was very diverse in terms of race and gender, which may increase generalizability; however, a larger (perhaps nationally representative) sampling design would provide better estimates of true sensitivity and specificity of these tools. Second, we did not have a non-college non-clinical community sample, which calls into question the external validity of these findings to dissimilar groups. Third, any estimate contains uncertainty, and the top two items in our sample may not be the top two at the population level (see Appendix for other candidate items with similarly strong properties). Fourth, although the brief yes-no format of responses improves the ease of use of the identified measure, more informative responses would be expected to

increase discriminative power. In particular, we believe that psychosis screeners should query for distress associated with endorsed symptoms, given that distress is likely a distinguishing feature between clinical and non-clinical psychosis (see Kline et al., 2014; van Os et al., 2009).

Overall, the present study suggests that very brief (two-item) and simple (yes-no) screens for psychosis may be a viable first step for psychosis screening, especially in settings with low base rates. Further research should attempt to replicate our results, test the validity of the identified screens in varied settings including general and clinical populations, and consider strategies for addressing barriers to implementation (e.g., supporting community providers on referral options for specialty care for individuals who screen positive) to determine its potential real-world public health impact.

Funding

This work was supported by the National Institute of Mental Health (grants R01MH112612 and R34MH110506 to J.S.) and the Maryland Department of Health and Mental Hygiene, Behavioral Health Administration through the Center for Excellence on Early Intervention for Serious Mental Illness (OPASS# 14-13717G/M00B4400241 to J.S.).

References

Furr, D.C, 2016. Hierarchical Two-Parameter Logistic Item Response Model. *Stan Case Studies*. Available at: http://mc-stan.org/users/documentation/case-studies/hierarchical_2pl.html.

Furr, D.C, Seung, Y.L., Lee, J., Rabe-Hesketh, S., 2016. Two-Parameter Logistic Item Response Model. *Stan Case Studies*. Available at: http://mc-stan.org/users/documentation/case-studies/tutorial_twopl.html.

Fusar-Poli, P., Schultze-Lutter, F., Cappucciati, M., Rutigliano, G., Bonoldi, I., Stahl, D., et al., 2015. The dark side of the moon: meta-analytical impact of recruitment strategies on risk enrichment in the clinical high risk state for psychosis. Schizophr. Bull. 42 (3), 732-743.

Fusar-Poli, P., Cappucciati, M., Rutigliano, G., Lee, T.Y., Beverly, Q., Bonoldi, I., et al., 2016. Towards a standard psychometric diagnostic interview for subjects at ultra high risk of psychosis: CAARMS versus SIPS. Psychiatry journal. 1-111.

Hinkin, T.R., 1998. A brief tutorial on the development of measures for use in survey questionnaires. Organ. Res. Methods. 1 (1), 104-121.

Kelleher, I., Harley, M., Murtagh, A. Cannon, M., 2011. Are screening instruments valid for psychotic-like experiences? A validation study of screening questions for psychotic-like experiences using in-depth clinical interview. Schizophr. Bull. 37 (2), 362-369.

Kessler, R.C., Barker, P.R., Colpe, L.J., Epstein, J.F., Gfroerer, J.C., Hiripi, E., et al., 2003. Screening for serious mental illness in the general population. Arch. Gen. Psychiatry, 60 (2), 184-189.

Kline, E., Thompson, E., Bussell, K., Pitts, S.C., Reeves, G. Schiffman, J., 2014. Psychosis-like experiences and distress among adolescents using mental health services. Schizophr. Res.152 (2), 498-502.

Kline, E., Wilson, C., Ereshefsky, S., Denenny, D., Thompson, E., Pitts, S.C., et al., 2012. Psychosis risk screening in youth: a validation study of three self-report measures of attenuated psychosis symptoms. Schizophr. Res. 141 (1), 72-77.

Kroenke, K., Spitzer, R.L. Williams, J.B., 2003. The Patient Health Questionnaire-2: validity of a two-item depression screener. Medical care. 41 (11), 1284-1292.

Linscott, R.J., Van Os, J., 2013. An updated and conservative systematic review and meta-analysis of epidemiological evidence on psychotic experiences in children and adults: on the pathway from proneness to persistence to dimensional expression across mental disorders. Psychol. Med. 43 (6), 1133-1149.

Loewy, R.L., Bearden, C.E., Johnson, J.K., Raine, A. Cannon, T.D., 2005. The prodromal questionnaire (PQ): preliminary validation of a self-report screening measure for prodromal and psychotic syndromes. Schizophr. Res. 79 (1), 117-125.

Loewy, R.L., Pearson, R., Vinogradov, S., Bearden, C.E. Cannon, T.D., 2011. Psychosis risk screening with the Prodromal Questionnaire—brief version (PQ-B). Schizophr. Res. 129 (1), 42-46.

Mead, A.D. Meade, A.W., 2010. Item selection using CTT and IRT with unrepresentative samples. In *Twenty-Fifth Annual Meeting of the Society for Industrial and Organizational Psychology*. Atlanta, GA.

Miller, T.J., McGlashan, T.H., Rosen, J.L., Cadenhead, K., Ventura, J., McFarlane, W., et al., 2003. Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. Schizophr. Bull. 29 (4), 703-715.

Miller, T.J., Cicchetti, D., Markovich, P.J., McGlashan, T.H., Woods, S.W. (2004). The SIPS screen: a brief self-report screen to detect the schizophrenia prodrome. Schizophr. Res. 70 (1), 78.

Millman, Z.B., Pitts, S.C., Thompson, E., Kline, E.R., Demro, C., Weintraub, M.J., et al., 2018. Perceived social stress and symptom severity among help-seeking adolescents with versus without clinical high-risk for psychosis. Schizophr. Res. 192, 364-370.

Ord, L., Myles-Worsley, M., Blailes, F., Ngiralmau, H., 2004. Screening for prodromal adolescents in an isolated high-risk population. Schizophr. Res. 71 (2–3), 507-508.

Savill, M., Skymba, H.V., Ragland, J.D., Niendam, T., Loewy, R.L., Lesh, T.A., et al., 2018. Acceptability of Psychosis Screening and Factors Affecting Its Implementation: Interviews With Community Health Care Providers. Psychiatr Serv. 69 (6), 689-695.

Schiffman, J., 2018. Considerations for the development and implementation of brief screening tools in the identification of early psychosis. Schizophr. Res. 199, 41-43.

Stan Development Team, 2017. RStan: the R interface to Stan. R package version 2.16.2.

Van Os, J., Linscott, R.J., Myin-Germeys, I., Delespaul, P. Krabbendam, L., 2009. A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness–persistence–impairment model of psychotic disorder. Psychol. Med. 39 (2), 179-195.

Appendix: Receiver Operating Characteristics for 4 two-item scales

The top five performing items in our sample were as follows:

Y-PARQ 12: Do you ever hear the voice of someone talking that other people cannot hear?
Y-PARQ 20: Do familiar surroundings sometimes seem unreal to you?
Y-PARQ 22: Do you see things that others can't or don't see?
Y-PARQ 28: Have you ever felt that someone was playing with your mind?
PQ-B 20: Have you seen things that other people can't see or don't seem to see?

Y-PARQ items 22 and 28 had the strongest properties in our sample. The other three items had >25% chances of outperforming Y-PARQ item 28 at the population level. No item on any screener had a >25% chance of outperforming Y-PARQ item 22. Therefore, we give receiver operating characteristics for five credible two-item scales.

Y-PARQ 22 and Y-PARQ 28

a. Operating characteristics of two-item screener in sample of 472 clinical and non-clinical individuals.

	Clinical High Risk or Psychotic Disorder (n=71)		Active hallucinations or delusions (n=36)	
Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	92%	59%	90%	78%
2	99%	32%	98%	53%

b. Operating characteristics of two-item screener in sample of 139 help-seeking individuals

Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	68%	59%	65%	78%
2	95%	32%	93%	53%

Y-PARQ 22 and Y-PARQ 12

a. Operating characteristics of two-item screener in sample of 473 clinical and non-clinical individuals.

	Clinical High Risk or Psychotic Disorder (n=71)		Active hallucinations or delusions (n=36)	
Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	93%	65%	90%	86%
2	98%	44%	96%	67%

b. Operating characteristics of two-item screener in sample of 139 help-seeking individuals

Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	73%	65%	68%	86%
2	87%	44%	84%	67%

Y-PARQ 22 and Y-PARQ 20

a. Operating characteristics of two-item screener in sample of 471 clinical and non-clinical individuals.

	Clinical High Risk or Psychotic Disorder (n=71)		Active hallucinations or delusions (n=36)	
Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	92%	59%	89%	69%
2	99%	25%	98%	42%

b. Operating characteristics of two-item screener in sample of 139 help-seeking individuals

Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	71%	59%	64%	69%
2	94%	25%	93%	42%

Y-PARQ 22 and PQ-B 20

a. Operating characteristics of two-item screener in sample of 473 clinical and non-clinical individuals.

	Clinical High Risk or Psychotic Disorder (n=71)		Active hallucinations or delusions (n=36)	
Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	87%	58%	86%	86%
2	97%	44%	97%	67%

b. Operating characteristics of two-item screener in sample of 139 help-seeking individuals

Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	70%	58%	71%	86%
2	82%	44%	81%	67%

Table 1

Participant characteristics

Cohort	
Clinical	139
College	335
Gender	
Males	213
Females	261
Age	
Mean	19
SD	3.5
Race	
White	181
Black	122
Asian	99
Multiethnic	50
Native American	2
Other	12
Did not report or unknown	8

Table 2

Table 2a. Operating characteristics of two-item screener in sample of 471 clinical and non-clinical individuals for two potential cut points.

	Clinical High Risk or Psychotic Disorder (n=71)		Active hallucinations or delusions (n=36)	
Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	92%	59%	90%	78%
2	99%	32%	98%	53%

Table 2b. Operating characteristics of two-item screener in sample of 139 help-seeking individuals

	Clinical High Risk or Psychotic Disorder (n=71)		Active hallucinations or delusions (n=36)	
Cutoff	Specificity	Sensitivity	Specificity	Sensitivity
1	68%	59%	65%	78%
2	95%	32%	93%	53%

Table 3

Correlations Among Two-Item Screen and Key Study Variables in College Sample

	Two-item screen	*Prime*	*PQ-B*	*YPARQ-B*	*SPQ*	*BDI-II*	*BAI-II*
Two-item screen	1	0.49	0.64	0.63	0.40	0.29	0.42
Prime		1	0.66	0.81	0.16	0.31	0.36
PQ-B			1	0.71	0.57	0.49	0.51
YPARQ-B				1	0.26	0.24	0.19
SPQ					1	0.52	0.54
BDI-II						1	0.54
BAI-II							1

Notes. PQ-B = Prodromal Questionnaire-Brief. YPARQ-B = Youth Psychosis at Risk Questionnaire-Brief. SQP = Schizotypal Personality Questionnaire. BDI-II = Beck Depression Inventory-II. BAI-II = Beck Anxiety Inventory-II. *N*'s range from 308 to 335 due to occasional missing data. All correlations significant at $p<.01$. All correlations between two-item screen and other variables significant at $p<.0001$.

Figure 1 .Estimated severity parameters for full sample, color coded according to the presence of active hallucinations or delusions.

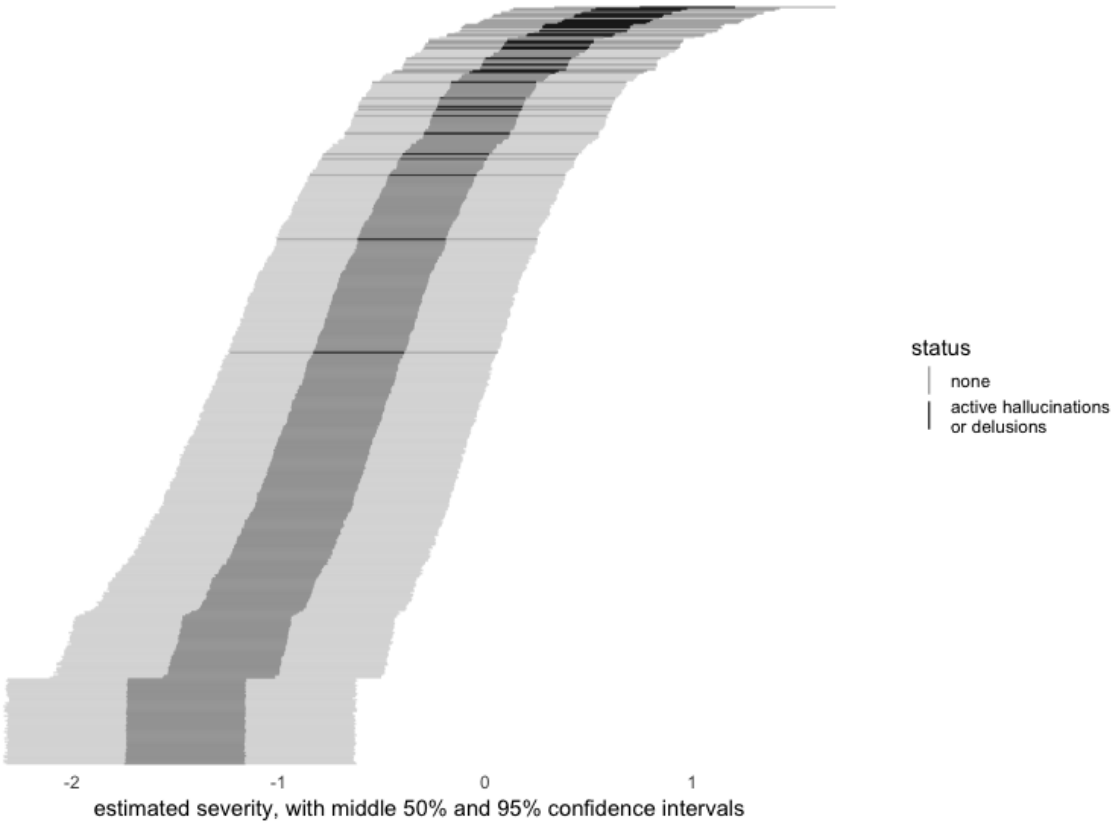

Figure 2 .Item information curves. X-axis gives latent psychosis severity expressed as a percentile with respect to the full sample. The peak of each curve is located at the level of severity at which an estimated 50% of people with that level of severity would endorse that item. Dotted line marks the median severity for people with active hallucinations or delusions (clinician-rated). Top five items at the criterion are plotted in black. In order of performance (from highest to lowest), these are: Y-PARQ 22, Y-PARQ 28, Y-PARQ 12, PQ-B 20, Y-PARQ 20.