Computerized Adaptive Testing (CAT) and the Future of Measurement-Based Mental Health Care

Andrew D. Carlo¹ · Brian S. Barnett² · David Cella³

Accepted: 16 February 2021 / Published online: 2 March 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Modern health care demands a foundation grounded in measurement that emphasizes patient-centricity. Consequently, innovations in outcome assessment and symptom quantification have become crucial pursuits in many fields of medicine, including mental health (MH).

In recent years, measurement-based care for MH disorders has reached a tipping point (Fortney et al. 2017) due to its association with improved clinical outcomes, as well as the abundance of publicly accessible scales, availability of digital administration modalities, and growing focus of payers and purchasers on value. Further, with robust data revealing that MH conditions impact outcomes for other medical problems, health care systems are increasingly expanding measurement-based care strategies for MH across various treatment settings.

Still, measurement-based care in MH faces formidable and unique challenges, ranging from ineffective clinician- and health system-level financial incentivization to a lack of user-friendly digital tools designed to streamline assessments. Perhaps most troublesome is the fact that nearly all MH outcomes are patient reported and require a commitment to longitudinal assessment. This has complicated efforts to achieve a high level of fidelity in MH outcome measurement, as repeated patient and staff input is challenging to sustain.

Solving the measurement-based care conundrum could not only revolutionize MH research and practice, but it could also facilitate continued integration into the larger health

care system. Therefore, we argue that it is time to advance measurement-based care in MH by replacing static measurements of illness severity with computerized adaptive testing (CAT), which has the potential to optimize large-scale outcome assessment, while also enhancing measurement fidelity over time (Gibbons et al. 2016; Kroenke et al. 2020).

The need for CAT is particularly acute for depression care, given this condition's high prevalence and substantial associated morbidity. CAT measurement of depression, as well as related patient-centered health domains (e.g., anxiety, sleep, social function), enables brief assessment that is also accurate at the individual level. Fifty years of testing, prior to the introduction of CAT, failed to deliver this now-possible “brief precision.”

Among validated depression rating scales, the Patient Health Questionnaire-9 (PHQ-9) has historically been one of the most widely employed. The PHQ-9 has a number of helpful attributes, as it: (1) is evidence-based across various settings and populations, (2) maps onto the Diagnostic and Statistical Manual of Mental Disorders (DSM), (3) has evidence-based score cut-offs (e.g., < 5 for remission), and (4) can be administered either synchronously or asynchronously with clinical staff. For these and other reasons, it was recently highlighted as a candidate instrument for harmonizing depression outcome measures in research and clinical practice (Gliklich et al. 2020).

However, the PHQ-9 and similar instruments are constrained by their administration inefficiency, static design, and dearth of patient-important components (Chevance et al. 2020). PHQ-9 question items do not change between administrations and each is weighted equally. Therefore, even if a patient has never noted sleep difficulties across multiple PHQ-9 administrations, they will continue to be presented with sleep-related questions each time. This rigid, repetitive approach is inefficient and may diminish engagement, thereby increasing the risk of response set bias as depression symptom severity changes over time (Gibbons et al. 2016). Additionally, the lack of error estimates in static instruments

✉ Andrew D. Carlo
andrew.carlo@nm.org

¹ Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, 446 E. Ontario St. #7-200, Chicago, IL 60611, USA

² Department of Psychiatry and Psychology, Center for Behavioral Health, Neurological Institute, Cleveland Clinic, Cleveland, OH, USA

³ Department of Medical Social Sciences, Northwestern Feinberg School of Medicine, Chicago, IL, USA

like the PHQ-9 precludes the determination of measurement certainty for a given patient.

Novel instruments incorporating CAT approach measurement-based care differently than classical instruments by leveraging item response theory (see Table 1 for a summary of noteworthy contrasts and similarities). Unlike classical instruments, those incorporating item response theory are built upon large question banks composed of items with varying severity level ratings (Cella et al. 2007). Higher-level questions assess more advanced levels of illness (e.g., severe functional impairment), while lower-level questions target the opposite (e.g., feeling sad occasionally). Since item response theory-based questions are ordered by level, patients can first be presented with a question targeting the median illness severity. Depending on initial response, CAT algorithms will tailor subsequent questions to the detected level of illness severity. Administration ceases after a certain number of items are completed or the standard error falls below a pre-determined threshold (Cella et al. 2007). This approach can improve efficiency, since, unlike in legacy instruments, there is no obligation to present the same number of items during each assessment. In fact, previous research demonstrates that CAT approaches reduce the total number of items administered by an average of 50%, with no reduction in measurement precision (Gibbons et al. 2016).

For example, the Patient-Reported Outcomes Measurement Information System Depression (PROMIS-D) measure

leverages CAT algorithms to present the fewest number of questions needed to obtain a precise depression symptom score during each administration. For adults, the minimum number of items administered is four, and the measure is stopped after either 12 items are administered or the standard error is below a pre-determined threshold (HealthMeasures, 2020). The median number of items per administration is four (Pilkonis et al. 2014). This can save valuable clinical time, particularly when tests are iteratively administered to large numbers of patients, with one recent article describing a real-world PROMIS-D implementation in a dermatology clinic requiring an administration time of only 1.1 min on average (compared to 2 min for the PHQ-9) (Gaufin et al. 2020).

CAT instruments are also inherently customizable, enabling the use of interchangeable items and personalization (Cella et al. 2007). Since the CAT algorithm incorporates data from previous responses, the instrument may differ during each administration, with patients always being presented with the most symptom-relevant questions. These attributes may promote clinician and patient engagement, while also allowing scores derived from different questions to be meaningfully compared on the same scale (Cella et al. 2007; Gibbons et al. 2016). Importantly, CAT instruments achieve a given level of measurement precision over successive administrations much more quickly than legacy

Table 1 Comparisons between adaptive and static testing

	Computerized adaptive testing	Legacy (static) testing
Common Example	Patient-Reported Outcomes Measurement Information System—Depression (PROMIS-D)	Patient Health Questionnaire-9 (PHQ-9)
Measurement Theory	Item Response Theory (IRT)	Classical Testing Theory (CTT)
Allows for Linking Across Depression Measurement Instruments	Yes	No
Customizable	Yes	No
Same Question Items Each Administration	No	Yes
Same Number of Questions Each Administration	No	Yes
Precision at the Individual Level	Yes	No
Inclusion of Questions about Suicidal Ideation	Sometimes (though these questions may be omitted)	Sometimes (though many include questions about suicidality)
Scoring Method	Digital	Digital or manual (depending on format of instrument)
Digital Administration Capability	Yes (required)	Sometimes (may be digital or non-digital)
Interoperability with the Electronic Medical Record	Yes (directly or through an application programming interface)	Sometimes (when digitally administered, may integrate directly or through an application programming interface)
Associated Financial Costs	Yes (cost of technology)	No (often publicly available)
Staff Administration Requirement	No (administered electronically)	Sometimes (may be administered by staff or self-administered)
Requires Staff Training	Yes	Yes

measures that administer the same items to all respondents (Gibbons et al. 2016).

Digitally administered CAT assessments can be completed anywhere at any time and on a variety of electronic devices, making them well suited for the COVID-19 era, and scores can be integrated into electronic medical records. Recently published studies have evaluated score cut-offs and the minimally important difference (i.e., the smallest clinically relevant score change) for CAT instruments such as the PROMIS-D, thereby enhancing their clinical utility (Kroenke et al. 2020). PROMIS-D and similar instruments also have established crosswalk linkage tables for legacy static depression outcome measures like the PHQ-9 (Choi et al. 2014; Gliklich et al. 2020), facilitating direct score comparisons and conversions.

Like all approaches, CAT has limitations. Due to computational demands, CAT instruments must be administered digitally, which may be a barrier in some settings. Additionally, as with any new health care technology, implementation of CAT may pose a financial cost to health systems, and patients, staff and clinicians must ascend a learning curve. However, we believe that these disadvantages are outweighed by the ability of CAT instruments to improve outcome measurement efficiency and precision.

Though measurement-based care is rapidly becoming essential to the optimal management of common MH problems, we have yet to capitalize on newer and more effective approaches to obtaining these measurements in both care delivery and research settings. CAT is validated, immediately actionable, and unhindered by many of the limitations of static instruments. Should we choose to embrace it, CAT has the potential to make measurement-based care part of everyday practice in the treatment of mental illness.

Author Contributions This manuscript has not been previously published and is not under consideration in the same or substantially similar form in any other peer-reviewed media. All authors listed have contributed sufficiently to the project to be included as authors, and all those who are qualified to be authors are listed in the author byline.

Funding The authors did not receive support from any organization for the submitted work.

Data Availability The submitted work does not contain any original data.

Declarations

Conflict of interest Dr. AC and Dr. BB report no conflicts of interest. Dr. DC reports NIH funding to his institution and a position as

an unpaid board member of PROMIS Health Organization, a 501(c)(3) foundation.

References

- Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care*, 45(Suppl1), S3–S4. <https://doi.org/10.1097/01.mlr.0000258615.42478.55>
- Chevance, A., Ravaud, P., Tomlinson, A., Le Berre, C., Teufer, B., Touboul, S., et al. (2020). Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: Qualitative content analysis of a large international online survey. *The Lancet Psychiatry*, 7(8), 692–702. [https://doi.org/10.1016/S2215-0366\(20\)30191-7](https://doi.org/10.1016/S2215-0366(20)30191-7)
- Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. *Psychological Assessment*, 26(2), 513–527. <https://doi.org/10.1037/a0035768>
- Fortney, J. C., Unützer, J., Wrenn, G., Pyne, J. M., Smith, G. R., Schoenbaum, M., et al. (2017). A tipping point for measurement-based care. *Psychiatric Services*, 68(2), 179–188. <https://doi.org/10.1176/appi.ps.201500439>
- Gaufin, M., Hess, R., Hopkins, Z. H., Biber, J. E., & Secrest, A. M. (2020). Practical screening for depression in dermatology: Using technology to improve care. *British Journal of Dermatology*, 182(3), 786–787. <https://doi.org/10.1111/bjd.18514>
- Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology*, 12, 83–104. <https://doi.org/10.1146/annurev-clinpsy-021815-093634>
- Gliklich, R. E., Leavy, M. B., Cosgrove, L., Simon, G. E., Gaynes, B. N., Peterson, L. E., et al. (2020). Harmonized outcome measures for use in depression patient registries and clinical practice. *Annals of Internal Medicine*, 172(12), 803–809. <https://doi.org/10.7326/M19-3818>
- HealthMeasures. (2020). *HealthMeasures—Transforming How Health is Measured*. Retrieved September 24, 2020, from <https://www.healthmeasures.net/resource-center/measurement-science/computer-adaptive-tests-cats>.
- Kroenke, K., Stump, T. E., Chen, C. X., Kean, J., Bair, M. J., Damush, T. M., et al. (2020). Minimally important differences and severity thresholds are estimated for the PROMIS depression scales from three randomized clinical trials. *Journal of Affective Disorders*, 266, 100–108. <https://doi.org/10.1016/j.jad.2020.01.101>
- Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C., & Lawrence, S. M. (2014). Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®) in a three-month observational study. *Journal of Psychiatric Research*, 56(1), 112–119. <https://doi.org/10.1016/j.jpsychires.2014.05.010>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.