Neuropsychology Review (2023) 33:628–642
https://doi.org/10.1007/s11065-023-09605-3

REVIEW

Invited Commentary: Advancing but not yet Advanced: Assessment of Effort/Malingering in Forensic and Clinical Settings

David Faust[1,2]

Received: 11 April 2023 / Accepted: 2 May 2023 / Published online: 18 August 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Neuropsychologists' conclusions and courtroom testimony on malingering can have profound impact. Intensive and ingenious research has advanced our capacities to identify both insufficient and sufficient effort and thus make worthy contributions to just conflict resolution. Nevertheless, given multiple converging factors, such as misleadingly high accuracy rates in many studies, practitioners may well develop inflated confidence in methods for evaluating effort/malingering. Considerable research shows that overconfidence often increases diagnostic and predictive error and may lead to fixed conclusions when caution is better advised. Leonhard's work thus performs an important service by alerting us to methodological considerations and shortcomings that can generate misimpressions about the efficacy of effort/malingering assessment. The present commentary covers various additional complicating factors in malingering assessment, including other factors that also inflate confidence; subtle and perhaps underappreciated methodological flaws that are inversely related to positive study outcomes (i.e., the worse the flaws the better methods appear to be); oversimplified classifications schemes for studying and evaluating effort that overlook, for example, common mixed presentations (e.g., malingering and genuinely injured); and the need to expand research across a greater range and severity of neuropsychological conditions and diverse groups. More generally, although endorsing various points that Leonhard raises, a number of questions and concerns are presented, such as methods for calculating the impact of case exclusions in studies. Ultimately, although Leonhard's conclusions may be more negative than is justified, it seems fair to categorize methods for assessing malingering/effort as advancing, but not yet advanced, with much more needed to be done to approach that latter status.

Keywords Malingering · Assessment of malingering · Assessment of effort · Neuropsychological assessment · Forensic · Neuropsychology and law · Malingering tests

I come to the issue of malingering/effort assessment in neuropsychology having performed an early study on the topic (Faust et al., 1988). I have also participated in numerous cases across decades as a consultant or testifying expert in defense and plaintiff cases in the civil (and occasionally criminal) arena. In many of these cases, possible malingering was a prominent, or at least secondary, consideration. Well beyond my personal observations, it is evident that the consequences of testimony on malingering can be profound. Although legal outcomes depend on an array of evidence and considerations, testimony on malingering can be decisive. A false-positive conclusion may result in a person desperately needing treatment being unable to afford it, or a false-negative error to discharging someone from a forensic facility who in truth has continuing, but unrecognized, plans to commit murder, with few restraints now remaining in place.

In the courtroom, claims about psychological and neuropsychological status, and particularly assessment of quality of life issues, often rest substantially on circumstantial evidence and the appraisal of hypothetical constructs (e.g., malingering, memory, depression), rather than directly observable physical injuries or events. Given the importance of such high-inference issues in many forensic cases and the nature of the subject matter, psychologists and neuropsychologists are often particularly suited and needed to assist

✉ David Faust
faust@uri.edu

[1] Department of Psychology, University of Rhode Island, 142 Flagg Rd., Kingston, RI 02881, USA

[2] Department of Psychiatry and Human Behavior, Warren Alpert Medical School of Brown University, Providence, RI, USA

in appraising litigants' status. Considered together with the significance of malingering/effort assessment in the legal arena, the explosion of research in this area has been a welcome development.

Research on malingering/effort assessment and the accuracy of self-report has often reflected innovative thought and ingenuity that have moved the field forward. A major aim of such efforts is to advance the normative goal of just conflict resolution by helping verify true injury or raise doubt about false claims. However, as Leonhard's (2023) two-part work suggests, error may subvert the process of effort assessment more often than we realize and potentially move us further from, rather than closer to, fulfilling our best intentions and pro-social aims.

Although I will raise questions about some of Leonhard's arguments or specific aspects of his positions, I believe there is considerable merit to most of the overriding concerns and issues he raises. I hope the field will take his critique with utmost seriousness and embrace the constructive and heuristic value of addressing key challenges that merit our best efforts and research creativity. Some of my colleagues seem to express the view that problems with malingering detection are largely resolved. Leonhard says it is not so. Is he overstating the case somewhat? Perhaps, but likely to a considerably lesser extent than some might assume. Possible overperception of scientific status may partly stem from what we might characterize as subtle complexities and methodological problems that, nonetheless, can have surprisingly robust, negative impacts. The progress that has been achieved in the appraisal of malingering/effort deserves praise, but the gap between *advancing* and *advanced* often remains substantial and might call for more reserved positions.

Before continuing, I think the tendency to sometimes refer to the sorts of measures under consideration here as malingering tests is unfortunate. Rather, for the most part, these tests and methods assess effort (with occasional exception, such as when someone performs far below chance level and there is good reason to believe they are intentionally underperforming or feigning deficit). All horses have hoofs, but all animals with hoofs are not horses. Clearly, some individuals might be unable to exert sufficient effort to obtain accurate measurement of their true or potential abilities, as might be the case with someone who is toxic on Tegretol (and, as paradoxical as it might sound, is doing as well as they can at that time). We would hardly want to identify such an individual as a malingerer based on insufficient effort alone.

Identifying questionable or insufficient effort can surely be of value. However, if one is trying to identify malingering, identifying insufficient effort, by itself, is rarely the final step in a sound decision process. As such, in my commentary, when I use the term *malingering*, I will often add a slash followed by a term like *effort* or *exaggeration*, or occasionally use one of these latter terms alone. Finally, when referring to what is often termed *symptom validity testing*, which Leonhard does address to an extent in his articles, I will instead usually use such descriptors as *response set* scales or simply refer to the accuracy of self-reports or informant's reports.

Methodological Flaws Often Converge to Inflate Accuracy Rates, with Practitioner Overconfidence Compounding This Problem

Design features and procedures common to research studies on the detection of malingering/effort tend to produce overestimations of accuracy, which may be substantial. As Leonhard covers extensively, one such procedure is excluding cases that are difficult to classify correctly, or so-called *too close to call* (TCTC) cases. A considerable percentage of cases may be excluded, approaching or exceeding 50%. This is like evaluating a quarterback's *overall* passing proficiency after eliminating the tougher throws. Obviously, excluding cases that are hard to classify leaves easier cases overall, or cases we are more capable of classifying correctly. The *extreme group problem* (Faust et al., 2021) is a closely related issue that stems from such exclusions and related practices, and which results in extreme study samples (those who are almost definitively or very likely malingering/making inadequate effort versus those who are almost definitively or very likely not malingering/or are making adequate effort). Hence, many of the less extreme or harder cases are eliminated, and study outcomes can markedly distort accuracy rates across cases as a whole.

Two things can make the extreme group problem and eliminating more ambiguous (TCTC) cases especially pernicious. First, the degree to which they distort or inflate accuracy rates often seems to be under-recognized. Second, there is a strong positive relationship between the extent of these methodological features, or what might be more aptly described as methodological shortcomings, and inflation in accuracy rates. All else being equal, the greater the methodological *problem*, the greater the *inflation* in accuracy rates or degree of misrepresentation, because the more one eliminates the harder cases, the more accuracy levels are artificially elevated.

There are elements of a worst case scenario here: methodological problems that may be hard to recognize but that, to the degree present, make methods look better than they are. The ultimate impact may not only be marked overestimations of accuracy, but distortion in the rank order of efficacy across methods. Keeping in mind that study methods can differ across malingering/effort tests or may only partly overlap (even in meta-analyses), seeming relative standing across tests may be determined as much, and sometimes more, by the degree of methodological flaw in the underlying studies than anything else (see further below, and Table 2 in Leonhard's second article). Thus, even proficient and highly conscientious practitioners may be drawn to methods that seem superior, but in fact are inferior, to alternative choices.

Other potential contributions to inflated accuracy rates can include the degree of redundancy among measures used in validation studies and the combination of effort tests practitioners select. Many studies evaluate the effectiveness of new methods for appraising malingering/effort by examining their agreement with other methods or measures, which causes what might be termed inter-correlational drift (Faust et al., 2021). The end result is often measures with much stronger associations (collinearity) with one another than may be assumed by researchers who perform Monte Carlo analyses, or who directly study combinations of measures, as opposed to combinations commonly used in practice or when clinicians create their own combinations of tests.

To illustrate, suppose a researcher uses results on the Test of Memory Malingering (TOMM) (Tombaugh, 1996) and a highly correlated measure to form criterion groups of malingerers versus non-malingerers, and then evaluates a new measure based on its agreement with assigned group membership. The higher the correlation between the new measure and the criterion measures, the better the new measure will look. (Ironically, this can make a new measure that improves on accuracy considerably appear flawed, because disagreements may seem to reflect errors the new measure makes.) A practitioner constructing a group of effort tests may well select a combination of measures that seem to have performed best in such validation studies. In contrast, Monte Carlo studies that assume lower collinearity among measures, or direct studies on combinations of tests that avoid highly inter-correlated measures, may generalize poorly to the practitioner's combination and greatly overestimate the surety of that practitioner's results. For example, failing, say, three of eight effort tests with relatively low or modest intercorrelations may yield strong evidence for malingering, but may markedly overestimate the likelihood of malingering when a practitioner uses a combination of eight tests with higher or much higher inter-correlations.

Speaking anecdotally, I have been involved in multiple cases in which defense experts argued that two or three failures on effort tests among a combination of five to ten measures they put together themselves suggested a 95%, or perhaps even a 99% + likelihood of malingering. One could see how such assertions originated in the expert's extrapolation of results from studies on test combinations that, however, assumed or used measures with considerably greater independence from one another than the expert's own combination. Experts who cite authors and publications and testify with genuine sincerity can sound highly credible, yet may have inadvertently fooled themselves from the onset. As a result, they may also end up fooling jurors, leading them to make decisions with potential catastrophic consequences for litigants. On the other side of the same coin, in some cases, plaintiff's experts who administered two or three highly correlated effort tests may have treated them as if they were largely independent measures, thereby opining that the results all but conclusively ruled out poor effort or falsification.

Inflated beliefs about the accuracy of effort assessment can compound what appears to be pervasive overconfidence among laypersons and professionals in multiple fields, mental health professionals included (Faust & Furman, 2022; Miller et al., 2015; Sieck & Arkes, 2005; Walfish et al., 2012). Various naturalistic conditions tend to inflate confidence, or lead practitioners to believe they are right more often than is actually the case. Such factors and conditions include selective feedback about accuracy, self-fulfilling prophecies, over-attention to or over-weighting confirming instances, and the reconstructive nature of recall. Experts may also under-attend to factors that suggest the need to adjust confidence levels downward when methods are generalized to new situations or applications. For example, experts might extend a study involving a certain combination of tests to a different set of measures they used, or studies involving college students might be extended too freely to individuals with widely contrasting sociodemographic backgrounds. Lawyers may also encourage experts to state their opinions with minimal reservation or strong conviction (fearing that more nuanced messages could be missed or misconstrued), thus potentially leading jurors to overestimate the strength of the expert's findings.

The extensive research on human judgment and decision making shows that overconfidence may be among the most potent forces in degrading judgmental accuracy (Faust et al., in press; Sieck & Arkes, 2005). Overconfident decision makers often draw conclusions too soon, fail to modify their views even when subsequent information should be convincing, let carelessness creep into decision processes, do not double-check their work, reach overly extreme conclusions (e.g., definitely malingering; definitely not malingering), fail to take corrective steps to enhance their accuracy, and countervail the output of the strongest available decision methods too freely and thereby routinely compromise their effectiveness. Thus, the combination of problematic research practices that often make malingering/effort methods look better than they are, and decision makers who are often more confident than they should be from the outset, can be especially detrimental to achieving just conflict resolution. Again, considering that testimony about malingering or cooperation with assessment procedures may influence decisions with major, and even life and death, consequences, the importance of the issues that Leonhard raises comes to the forefront.

The Scope of Leonhard's Critique Is Impressive, yet Additional Critical Issues Increase Complexity Substantially

Leonhard addresses a number of central issues in malingering/effort assessment, and it is apparent how much thought and effort went into his critique and analyses. Despite the broad scope of Leonhard's work, and arguably by necessity, it generally touches lightly on the overall complexity of the subject matter. When one considers some of these additional issues, the breadth and difficulties of the challenges before us highlight the premature nature of the suggestion by some that we have achieved advanced status or solved most of the weighty problems in the evaluation of malingering/effort.

For example, determining how to calculate posterior odds or calculate incremental validity raises vexing questions in effort assessment (and other domains), as Jewsbury's (2023) impressive commentary sets out so thoughtfully. Evaluating or refining possible solutions is especially challenging when one lacks reasonably strong reference standards or criteria for determining when and how often classifications are correct. For example, in principle, to optimize incremental validity, we seek variables that are maximally valid and minimally redundant. However, how well can we evaluate those properties if, in a considerable percentage of cases, we cannot determine whether someone is malingering with sufficient accuracy? Variables with perfect validity should correlate perfectly with one another (also eliminating the need to use more than one variable), but in the real world, it may be hard to determine how much of the association between variables reflects overlap in accurate versus erroneous measurement. Furthermore, studying the cases we know how to classify will do little or nothing to help us learn how to classify the cases we do not know how to classify at present, and such an approach has considerable potential to lead us down blind alleys or degrade our efficacy (more on this matter shortly).

This is not to suggest we need infallible criteria or gold standards to make progress. Many scientific endeavors begin with valid, but fallible, criteria that are sufficient to get the process moving forward. Additionally, if perfect or near-perfect criteria or methods for verifying malingering and non-malingering existed, and if they could be accessed and used without exceptional effort or cost, we would be using them. The availability of nearly flawless methods would likely reduce much of the need to conduct the hundreds or thousands of additional research studies on effort we can anticipate occurring in the future, although it still might be worth investigating such matters as potential means for reducing cost and increasing efficiency, or performing updates that may be required over time.

A Sampling of Areas Needing Further Research, and Conceptual, Methodological, and Theoretical Advances

What are some areas in the assessment of malingering, effort, cooperation with evaluation procedures, and the accuracy of self-reports and informant input that might be priorities moving forward? I am not suggesting Leonhard should have covered any of the subjects I will describe, but only that various such areas intersect with issues he raised. The list that follows is certainly incomplete, but I think a good argument could be made for the relevance of each item. As wide-ranging and broad as Leonhard's two articles are, they are, to some extent and understandably, relative exercises in narrowing and simplifying problems. Most of the items that follow serve to illustrate the breadth of relevant scientific problems facing us, and how ongoing research in such areas may help us distinguish when simplifications are relatively harmless and when, in contrast, they come at the cost of too much distortion.

Limits of Dichotomous Classification, Need for Greater Differentiation, Combined Presentations

As Leonhard noted, dichotomizing malingering and non-malingering is an oversimplification, but one he adopted broadly because many studies target this distinction and certain key statistics for examining efficacy require dichotomous classification. I would argue that this particular oversimplification, whether justified or not given the current state of the science, creates a number of ancillary problems. For example, it is often important to determine the degree and breadth of reduced effort. Suppose a plaintiff with an array of genuine neurocognitive deficits that impair major life functions, but seemingly without objective evidence of injury, exerts insufficient effort on memory testing. Perhaps, this individual did so to increase compensation, but perhaps instead true memory decline was so demoralizing that memory tasks had become extremely distressing and quickly led to disengagement with such activities. Furthermore, as is common, suppose the practitioner mainly used effort tests emphasizing memory. Applying a dichotomous classification system and perhaps utilizing publications suggesting that failure on two or more malingering/effort tests indicates a high probability of malingering, the expert concludes confidently, and testifies compellingly, that the plaintiff was falsifying impairment. The end result might be for jurors to overlook or reject most or all of the plaintiff's claims as legitimate and view them instead as other instances of fabrication.

At one point, I asked about 20 consecutive lawyers, with about a 70 to 30% split between defense and plaintiff's attorneys, what they believed was the most common presentation in head injury cases from among four possible choices: not injured *and* embellishing/exaggerating; injured *and not* embellishing/exaggerating; not injured *and* not embellishing/exaggerating; and injured *and* embellishing/exaggerating. Almost all answered injured *and* embellishing/exaggerating.

It is a virtual certainty that mixed presentations occur, and such cases may be relatively frequent or common. For example, various features of the legal system can systematically foster embellishment of true injury, such as long delays in obtaining desperately needed and just compensation that embitters litigants; lawyers who may encourage litigants to make the strongest possible cases for themselves; or litigants' skepticism about fair treatment from companies or defense experts that may lead them to believe they must over-represent problems in order to receive their just due.

Suppose hybrid presentations, such as injured and exaggerating (I + /M +), occur regularly. I am certainly not arguing that individuals should be rewarded for feigning deficit, and one can also understand why the legal system might wish to foster outcomes that extract some cost for falsification, even should legitimate injury be present, in order to deter such practices. However, unless one takes the position that any degree of falsification should lead to the total denial of compensation, even if the presentation, for example, is 95% genuine and 5% falsified (which we might glibly label *The Mother Teresa Requirement*), there is an important issue here that has been relatively neglected in scientific efforts.

Rather than dividing classifications into two categories (malingering versus not malingering) with four possible outcomes (valid-positive, false-positive, valid-negative, and false-negative judgments), we might need at least four categories, i.e., I + /M + (genuinely injured and malingering/embellishing), I + /M − (genuinely injured and not malingering/embellishing), I − /M + (not injured and malingering/embellishing), and I − /M − (not injured and not malingering/embellishing). Such categories would create 16 possible outcomes, e.g., in the case of I + /M +: valid-positive and valid-positive judgments, or classifying an individual who is genuinely injured and is also malingering/embellishing as genuinely injured and malingering/embellishing; valid-positive and false-negative judgments, or classifying an individual who is genuinely injured and malingering/embellishing as genuinely injured but not malingering/embellishing; and so on for the other two possibilities (false-negative and valid-positive judgments; and false-negative and false-negative judgments). Now, with four different categories and 16 possible judgment outcomes, things have suddenly become more complicated. Perhaps it is a little difficult to wrap one's mind around the possibility of simultaneous valid-positive and false-positive judgments and other such combinations, but again, these sorts of outcomes almost surely occur and can vary widely and substantively from only malingering/embellishing, or only not malingering/embellishing.

Complexity builds further if one adds matters of degree and breadth to each of these classifications, dimensions that could contribute substantially to helping carve nature at the joints in this domain. In addition, I have not even mentioned identification of factors that may lead to reduced effort, some of which flow from genuine injury (e.g., apathy or inertia associated with frontal injury) and some of which do not. Thus, simple dichotomous classification schemes, such as malingering and not malingering, are often not only rather crude beginning points, but simplifications that may lead jurors to draw overly general conclusions that may do more to create misunderstanding than clarity. Given that overly general and simplified judgments in appraising critical aspects of effort can have such deleterious consequences, it seems prudent to treat these more complex issues and questions as priorities as opposed to scientific orphans. At minimum, and surely some testifying experts proceed just in this way, we might openly acknowledge limits in our knowledge and areas in which we presently have insufficient scientific foundations to form trustworthy opinions.

Tests of Malingering or Effort? Sources of Compromised Effort and Inaccurate Self-Reports

Malingering is a loaded term that carries (as the philosopher might say) considerable surplus meaning, is inextricably tied to intentionality (at least in many instances), and is a hypothetical construct that cannot be reduced to a series of observations but also requires inference or inductive reasoning. (Space limitations necessitate truncated discussion of these issues, and I hope these statements do not sound overly declarative; for more detailed coverage and explanation, see Faust et al., 2021.) Labeling a measure a malingering test or referring to measures as such is questionable, because with occasional exception (e.g., results far below chance level), these measures are usually designed to (and do) assess effort.

Attempts to create a false impression of deficit, or deficit associated with an event, come in diverse forms and combinations. Some variations include slowing of responses, purposely producing erroneous responses, presenting false symptoms, exaggerating symptoms, false attribution (e.g., self-reports indicating that deficits were caused by the accident in question, as opposed to other events or behaviors, such as a chronic history of substance misuse), over-reporting level of past functioning and under-reporting past problems, and under-reporting positive functioning and over-reporting negative functioning when describing the post-accident period. In addition, effort can be diminished by a range of factors beside the intention to portray deficit.

Such alternative factors have the potential to lower scores on at least some performance validity tests, and particularly on embedded measures designed to detect results below expected levels for the injury or event in question. Even relatively modest declines in performance that are not due to malingering or intentional misrepresentation can elevate false-positive error rates. In some cases, the impact of alternative factors on efforts tests, such as serious psychological disorders, has been carefully studied, but in other cases, there is little or no research. The potential impact of alternative factors on so-called symptom validity tests or response set scales (i.e., scales designed to assess possible tendencies to under- or over-report symptoms) also merits brief mention.

Inaccuracies in self-report, or in attribution of causes for symptoms, might simply result from error rather than efforts to mislead. Suppose a physician at a busy Emergency Room evaluates someone shortly after a car accident. The physician mistakes such symptoms as headache and dizziness, which are actually secondary to neck strain and anxiety, as indicative of concussion. The medical professional lists the diagnosis in the medical records and conveys it to the patient, who naturally accepts the conclusion. If the injured individual, and later plaintiff, conveys this mistaken diagnosis to other practitioners in good faith, thereby passing the error forward unknowingly, this could hardly be considered malingering. Psychological status, such as severe affective disorder and PTSD, can also result in distorted perceptions and inaccurate self-reports about functional capacities or cognitive status. Such psychological disorders may also reduce performance on cognitive tests, and brain injuries associated with apathy or marked fatigue may lead to reduced effort or endurance, especially given the length of batteries some neuropsychologists administer in a single day. Well-designed validity tests may help sort out alternative causal factors, but also may be insufficiently developed to help much with such differentiations, or may be ineffective for such purposes.

The belief that practitioners can sort out these varying possible causes for results on validity tests or measures using clinical or impressionistic judgment should not be taken for granted. Formal evidence is needed to evaluate the accuracy of such beliefs. It may be sobering to reflect on common beliefs held not too long ago about the ability to identify malingering through clinical judgment and without the aid of specially designed methods, beliefs which now seem naïve. In addition, extensive literature on the ability of individuals to integrate complex data, professionals in the mental health field and other disciplines included, shows bounded limits, inefficiencies, and often unsatisfactory accuracy rates without the benefit of formal decision aids (see Dawes et al., 1989; Faust & Ahern, 2012; Faust et al., in press).

Given the range of factors that can alter scores on tests designed to assess performance validity, it might be advisable to refer to these measures as assessing effort, both in studies and in forensic contexts. Considering such revised language is not a matter of semantics but of substance, and it might be especially called for on most embedded or standard tests used for that purpose. Embedded tests or methods often do not create that much separation between individuals who are intentionally underperforming versus other groups, e.g., individuals with lower initial baselines, unexpectedly poor outcomes from injuries, more severe injuries than individuals often included in background studies, and those with prior conditions that interact adversely with newly sustained brain trauma.

For methods designed to assess the fidelity of responses on psychological measures or rating scales, whether the term *symptom validity* measures is better or worse than a term like *response set* measures or some other equivalent that connotes accuracy seems worth considering. Using the word *validity*, both when referencing the assessment of response sets and more generally when addressing properties of measures, has caused confusion for decades. Furthermore, *symptom validity* now has been partly conflated with malingering when used side by side with the term *performance validity tests*.

Suspect results or clear failures on malingering/effort tests, with the exception of definitive or nearly definitive evidence of malingering (e.g., far below chance performance), can be seen as an important step in identifying inadequate effort, but more as a beginning versus endpoint in the diagnostic process. In a fair number of cases, identifying deficient effort that raises serious concerns about the accuracy of results on standard neuropsychological tests may be about as far as the testifying expert can and should go because that is as far as the science can take us in the case at hand. In other cases, and quite possibly a fair percentage of the time, experts might state that there are indications that testing may under-represent true capacities, but that the results are inconclusive, or that they simply cannot determine whether effort was or was not sufficient. (Of course, conversely, experts may be fully justified in indicating that they evaluated effort and results suggested, or strongly suggested, that no problems with effort were indicated, or, depending on the measures that were used and their ceilings, that effort was satisfactory or positive.)

If one considers the percentage of TCTC cases in background studies, experts who always seem to arrive at strong conclusions about poor effort or excellent effort would seem potentially suspect. Experts might state that in comparison to study results, obtaining more information in the instant case may allow them to go further, which could be true, but there is often insufficient research to test such assertions. Again, there is an abundance of decision research showing that additional information may not increase accuracy, or may diminish accuracy when weaker or poor information nevertheless influences thinking, perhaps unintentionally;

and that attempts to integrate large data sets primarily on the basis of clinical or subjective judgment are considerably less efficacious than we might think (see Faust & Ahern, 2012; Faust et al., 2021, in press; Sieck & Arkes, 2005).

What Can We Conclude from Failed Effort Tests?

In many cases, our knowledge of specific relationships between reduced performance on effort tests and performance on standard neuropsychological tests is limited. There is of course the commonsense notion that failures on effort tests translate to suppressed scores on other measures, with evidence to support these beliefs (e.g., Bhowmick et al., 2021; Clark et al., 2014; Sherry et al., 2022). Such conclusions seem particularly warranted when failure occurs clearly and uniformly, or nearly so, across multiple other effort tests disseminated by content area and position/timing across a test battery.

In contrast, take, for example, a situation in which an individual with a moderate head injury fails three of ten effort measures, two of which fall slightly below cutoff points based primarily on investigations involving individuals with mild head injuries. The practitioner might extrapolate from studies examining the number or proportion of failures across batteries with a similar number of effort measures (although the measures in the studies may only overlap partly or minimally with the measures the practitioner used). Should we label this individual as malingering, and what are we to make of the seven measures on which the individual scored normally?

Perhaps the failed measures have a common content area (e.g., short-term verbal memory). Does this allow, or to what extent does it allow, extrapolation to performance in other areas across the overall battery that was administered, or do the negative results across the other effort tests suggest that results in other areas on the standard battery can be treated as likely valid? Were the failures due to the content area, or simply due to tests that are more successful in creating the appearance that items are harder than they are? We may be largely reduced to guessing if the scientific knowledge we need to answer such questions is lacking. If we assume we can be confident in our clinical or impressionistic judgment, a critic might ask whether a major rationale for many of the studies on the evaluation of effort is the evidence that our subjective appraisals in this domain are not as accurate as we would like. And if professionals have trouble with initial or more basic appraisal of effort based primarily on clinical judgment and recognize the benefits of formal effort tests backed by science, which I believe has increasingly become a broadly accepted view, what basis do we have to assume that considerably more complex judgments about effort and its specific impacts on testing results can achieve sufficient efficacy without the help of sound scientific foundations?

Need to Expand Studies of Conditions Other than Head Injury; Co-occurring Conditions and Possible Impacts; and Diverse Groups and Test Bias

Moving on to other complicating factors, the great bulk of studies on performance validity involve head injury, or more specifically mild head injury and concussion, which accords with their high rates of occurrence and frequency in legal cases. Nevertheless, it would be helpful if more research were available on effort assessment and symptom validity appraisal with other conditions one sees fairly often in the courtroom and in clinical practice, such as exposure to toxins (e.g., CO and lead), electrical injury, conditions leading to hypoxia and anoxia, PTSD, sepsis and various infectious diseases, and hormonal disorders, which may co-occur with head injuries. Neuropsychological functions that tend to be diminished or preserved across areas and conditions may well call for changes in interpretive methods or algorithms, as might variations in the magnitude of diminished functions across different conditions and the severity of those conditions. For example, some individuals with high level CO exposure suffer such extreme memory impairments that they may fail multiple effort tests focusing on memory (as so many effort measures do), especially those involving more than brief delays. It is disturbing to think about the potential consequences of misidentifying an individual with a devastating memory disorder as a malingerer.

In addition, in many courtroom cases in which the occurrence of brain injury is in question, individuals may have experienced other injuries and impairments (e.g., peripheral nervous system injuries, visual disorders, chronic and severe headaches, chronic pain), which can complicate appraisal of effort. For example, peripheral injuries may diminish motor capabilities and result in performances suggesting poor effort on such tests as those measuring finger tapping speed. Other conditions, such as severe PTSD, may interact with brain injuries and thereby diminish performance to levels below that expected for brain injury alone and thereby suggest poor effort, as might occur on an embedded test like Digit Span. Especially for measures or methods that look for performance well below expected levels (in cases of mild head injuries), the combination of head injury and prior conditions, or the co-occurrence of new conditions due to the event in question, may combine or interact to make true impairments seem feigned or exaggerated. Alternatively, co-occurring conditions that actually have minimal impact on effort tests may be misperceived as explanations for depressed results, when such performances truly are indicative of poor effort.

Methods for detecting problem effort that emphasize performance below expected levels may lead to many false-positive identifications among individuals with low initial baselines. For example, suppose a review article suggests that a scaled score of 5 or lower on the Digit Span subtest from the Wechsler Adult Intelligence Scale-IV (Weschler, 2008) raises concern about inadequate effort. Take an individual, however, with a low baseline, say a Full Scale IQ (FSIQ) score of 80, which falls 1.33 standard deviations below the mean. A comparable level of performance on the Digit Span subtest is a scaled score of 6. Hence, a Digit Span score of 5, or just 1/3rd of a standard deviation below a 6, could lead to a classification of poor effort. Given normal subtest scatter, obtaining a score 1/3rd of a standard deviation below an individual's overall baseline is far from unusual. In contrast, for an individual with a FSIQ score of 100, a scaled score of 5 falls 1 2/3rds standard deviations below that FSIQ level.

Test performances may be reduced by test bias or biasing factors, especially among diverse groups. If test bias lowers a score by a standard deviation or more, rates of false-positive identification may increase markedly and result in specificity levels well below commonly advocated professional standards. In addition, one should not expect uniform biasing effects within diverse groups (especially those identified or grouped through dubious indicators). Rather, the influence of test bias within those groups is likely to vary, and thus among those for whom biasing impacts are greater, false-positive rates are likely to increase, perhaps substantially. Identifying and measuring test bias in the nuanced manner needed here is not only a difficult problem, but one that has been minimally studied in effort assessment (although far more information is available on some response set measures).

Level of biasing effects is also likely to vary across tests, although not necessarily in the ways sometimes assumed, e.g., that verbal tests are influenced considerably more than nonverbal tests (in contrast, see, for example, Hambleton et al., 2004; Heaton et al., 2009). Neuropsychologists may prioritize test patterns when evaluating possible malingering or poor effort, approaches that may well be idiosyncratic and lack sufficient scientific investigation and foundation in the first place. Such problems are likely to be compounded if the degree of bias differs across tests from negligible to considerable, and more so in areas in which performance may be systematically stronger among diverse groups (e.g., Mulenga et al., 2001). Given within group and within test variation in biasing effects, and changes in directionality in some instances, the usual meaning of test results and especially attempts at pattern analysis can easily go far astray. More generally, because the direction of test bias is usually to diminish scores, the most likely result is to elevate the false-positive error rate and potentially cause systemic disadvantage or harm, in direct opposition to principles of fairness and equity that experts themselves may hold dear.

Single Base Rates Do Not Fit All, and the Need for Base Rates Extends Beyond Dichotomous Classifications

Regularly setting base rates for malingering/poor cooperation at around 0.40 for forensic populations and 0.10 for clinical populations is a questionable practice that does not optimize decision accuracy and likely can be improved over time. Leonhard adopted those levels for his current articles, presumably in part given the frequency with which they appear in the literature. Agreement among psychologists based on such sources as studies and observations from their practices is a sensible starting point for estimating base rates. However, aside from wide variations one may see in estimates, even when estimates tend to converge, they can be systematically influenced or distorted by a number of factors. For example, if our methods for identifying deficient versus sufficient effort are not necessarily as accurate as we think, and we generally lack definitive or near-definitive feedback about the accuracy of our judgments in this area, how can we determine whether our base rate estimates are sound?

As is well established, base rates can have a sizeable impact on diagnostic and predictive accuracy, and, not uncommonly, a base rate is the single strongest predictive variable. To illustrate, if something occurs (or does not occur) 90% of the time, then diagnostic signs and indicators must be more than 90% accurate to beat conclusions founded on base rates alone, e.g., if one always guesses a condition is present if it occurs 90% of the time in the setting of interest. (I am putting aside the potential differential impact of false-positive versus false-negative errors, which extends beyond the scope of my commentary.) Alternatively, suppose diagnostic signs and indicators point ten times more strongly to an infrequent neurodegenerative disorder versus Alzheimer's disease. Nonetheless, if Alzheimer's disease occurs 20 times more often than the alternative diagnostic condition in the setting of interest, then the differing base rates overwhelm the diagnostic indicators and Alzheimer's disease is the more likely choice. In many circumstances, one does not have to select between diagnostic indicators versus base rates but can combine the two to arrive at a more accurate estimate of likelihood. Very high or very low base rates, in particular, can have a huge impact on those likelihoods.

Not all groups in all clinical or forensic settings are likely to have similar rates of malingering or poor effort. For example, criminal defendants who seemingly have exhausted possible defenses, have no history of mental health disorder, but are now pursuing an insanity plea, likely have considerably lower base rates of major mental disorder than defendants who were hospitalized multiple times (including involuntarily) before the criminal offense occurred, and on each occasion were diagnosed with severe mental disorder. It also follows that the former group likely has a considerably higher base rate for malingering psychological disorder than the latter group. Alternatively, the base rate for faking bad in child custody cases is likely very low, and likely considerably lower than individuals who submitted claims for disability income following major injuries, were turned down for no apparent reason, and are now resubmitting claims.

Even if groups share some important common feature, such as clinical versus forensic status, if they differ in other important respects, then base rates may change considerably in relation to standing on these other features or variables. The rule of thumb is to select the narrowest applicable base rates, with narrowness conceptualized as variables that meet two criteria: (a) they alter base rates, and (b) they apply or are relevant to the individual of interest. For example, base rates may vary with age, setting or purpose of the evaluation (e.g., to name an obvious one, custody dispute versus personal injury claim), or the presence or absence of certain red flags for malingering or poor cooperation with assessment procedures. However, some of these variables may not be applicable to the examinee, such as differences in base rates between individuals who are 30–40 years of age versus 40–50 years of age when the examinee is 15 years old.

Base rates are a must to determine the effectiveness of diagnostic or predictive measures and variables in settings of application. One seeks base rates that are as accurate as possible, or at least reasonable approximations. If many cases in studies are TCTC, then should we expect such cases to occur much less often in practice (especially considering that many of the individuals in studies are drawn from forensic samples)? Suppose, however, that base rate estimates come primarily from cases that are clear enough to call but exclude more ambiguous or difficult cases. In such circumstances, one only has base rates for a subgroup of the overall population, which in turn may only generalize reasonably well to other relatively clear or obvious cases. Clear or obvious cases are the ones for which we least need help. Whether base rates with clear cases can be extrapolated to ambiguous cases is uncertain and may well be subject to considerable error. If base rate estimates include more ambiguous or difficult cases, then the error rate in classifying these cases correctly is in question from the start, and resultant base rate estimates may vary widely and have a large margin of error.

Substantial differences in accuracy can result when base rates differ markedly. Sometimes, a reasonable range for base rates can be determined, and in those instances, one should be able to create error terms that take that range into account. At other times, however, base rate estimates are so tenuous and vary so widely that insufficient information is available to appraise level of effort or cooperation with a reasonable degree of certainty, especially when measures produce something other than highly robust results. There is no shame in such open admissions, despite possible pressure from lawyers, or an opposing expert who nevertheless feels unconstrained in making declarative statements. It may be left to the more prudent expert to explain why confident conclusions about effort in the present case are likely illusory and exceed the bounds of knowledge.

In addition, if we adopt more complex classification schemes, such as one that includes the dual dimensions of injury and effort, the limitations of current base rates become all the more apparent. Taking other dimensions into account that seem to be of considerable importance, such as the level and breadth of noncooperation, adds further complexity and challenges to determining base rates and maximizing their utility.

There are a number of methods for estimating base rates in a range of ambiguous situations (e.g., see Jewsbury & Bowden, 2014; Faust et al., 2021). However, in substantial part, knowledge of base rates will grow in tandem with the gradual development of improved methods for identifying compromised effort given the dialectical relationship between the two. This is much like the situation with scientific classification and definition: advances in classification and definition are as much a product of advances in scientific knowledge, as advancement in scientific knowledge is a product of better classification and definition.

Additional Complicating Problems and Research Needs

Some other complicating problems and needs that seem worthy of emphasis include aspects of symptom validity testing; informants and informant questionnaires; defensiveness or dissimulation, i.e., “faking good”; and coaching, preparation, and transparency. The standards for developing symptom validity scales are sometimes shockingly lax. Various psychological tests and questionnaires include symptom validity scales for which there is little or no peer-reviewed literature, an obvious concern given the potential impact of results on such scales in shaping important decisions. Lack of adequate validation seems to be more the rule than the exception for questionnaires and rating scales designed for informants. Informants, especially if selected carefully for both their familiarity with a litigant and potential neutrality, can provide critical information for cross-checking the litigant’s self-reports and other informational sources, such as medical or mental health records. Diagnoses and conclusions in medical and mental health records may rest primarily on the history the litigant provided, especially when hard evidence for injury is lacking, as is common with such conditions as concussion. Alternatively, some individuals with brain injuries lack insight and understate problems substantially, something that can become apparent by obtaining informants’ reports.

Although my observational base may be skewed, for what it may be worth, I have been involved in numerous cases in which experts use response set measures, especially those designed for informants, with obviously deficient scientific foundations. A number of these response set scales have never been subjected to a single formal study. Manuals may fail to emphasize the very tenuous nature of such scales, and ironically, there is often little genuine or effective oversight for the use of such measures. Our ethical standards commonly emphasize the need to carefully scrutinize scientific evidence for the procedures we use, or advocate for limiting practice to the bounds of our competence. However, how can one practice competently when applying measures or working in domains in which subjective judgment has been found to be lacking (a primary rationale for creating properly validated measures of effort and cooperation) and virtually no science exists?

Detecting dissimulation, or faking good, often poses particular problems. There is far more research on faking bad than faking good, and clearly a limitation in the number of dissimulation scales with solid scientific backing. Although dissimulation often takes a second seat to interest in, and research on, malingering/exaggeration, it does come up in many cases. Examples include evaluations for parole programs, related evaluations for sociopathic characteristics, child custody evaluations, and increasingly in evaluations for return to work or sports with individuals who are highly motivated to resume activities, perhaps unwisely. Athletes that wish to return to sports as soon as possible, including youth who may be short-sighted or lack perspective on potential long-term costs, or professional athletes with enormous financial incentives, may commonly under-report symptoms. (Ironically, athletes may also purposely underperform on baseline cognitive testing so that, even if compromised, they can approximate their prior testing levels.) As concerning as various issues are that Leonhard raised in his critique of measures examining insufficient effort or exaggeration, one wonders what other important insights might result if he turned his attention to dissimulation measures.

In the Internet age, studies on coaching and the transparency of effort tests need to be extended. No matter how good effort tests may be or become, if their underlying designs become well understood and the basis for evolving strategies to beat them, many or most such tests will likely be susceptible to manipulation. Given similarities in the basic blueprint for many effort tests (e.g., designed to appear harder than is actually the case; very few errors expected despite their appearance to ensnare underperformers), even relatively narrow knowledge can assist in evading detection efforts, especially if practitioners select a few measures with similar fundamentals.

Coaching studies often provide overly general information, and it would be useful to conduct further investigations in which detection strategies and potential ways to beat measures are described explicitly and as completely as necessary. Information about measures and their design can often be found on the Internet, or may appear in books available at various websites, along with possible ways to avoid detection. Key details may also be described in publishers' advertisements or product descriptions that are not too difficult to access, and which may include pictures of manuals and related test materials. Lawyers who specialize in head injury cases may also be very familiar with a range of effort tests. In one study, Kovach (2017) provided brief, but explicit information on the TOMM and how to recognize the measure by sight (based on pictures found within seconds on the Internet). She then instructed individuals to feign deficit in as convincing a manner as possible. A considerable percentage of research participants produced abnormal scores on true memory tests (likely identifying them correctly), and everyone who did so escaped detection on the TOMM, yielding a false-negative rate of 100%.

Effort tests susceptible to explicit knowledge of design and strategies to circumvent them are likely to have short half-lives, and extensive research efforts may be compromised if the value of tests degrades rapidly and renders research on such measures largely obsolete. Adding to this problem, it may be hard to identify whether individuals have sufficient knowledge to beat tests and how widespread such knowledge might be, and thus to determine whether and to what extent false-negative error rates are increasing. Faust et al. (2021) provide a number of suggestions for decreasing susceptibility to knowledge of design and potential strategies to beat measures, such as varying the number of response options randomly, thereby capitalizing on the limits of the human mind to track multidimensional problems or dimensions simultaneously in real time.

Further Thoughts on TCTC Cases and the Extreme Group Problem

I wish to return to the TCTC and extreme group problem, and the important service Leonhard has performed by emphasizing the impact of high exclusion rates across numerous studies. If we seek to push the boundaries of knowledge, it is not the easy-to-classify cases, but the hard-to-classify cases, that are potentially of greatest interest. I do not hold as negative a view as Leonhard on the overall state of knowledge (see further below). In my view, various problems plaguing research on effort assessment are relatively common in the soft sciences when we direct concentrated efforts to challenging problems. Gradual, grinding progress with many starts and stops is far more the rule than the exception in everyday scientific life, and not the highly publicized (and sometimes overstated) leaps in knowledge.

Speaking broadly, I believe that ongoing advancements are reflected in the growing percentage of cases that we can classify with reasonable or even relatively high levels of certainty as demonstrating either sufficient or insufficient effort to form plausible conclusions about true level of functioning. Exactly what this percentage of cases might be cannot be determined with exactitude, and whether it is 10%, 20%, 30%, 40%, or some other figure is not something I will enter into here, except to address the following two points or opinions. First, that figure is likely to be higher, or considerably higher, than where we stood decades ago before the explosion of research on effort appraisal. Second, however, I believe that figure is lower than many experts assume, given (a) the types of critical problems that Leonhard (2023) has identified in his current work, such as the common exclusion of more difficult cases in background studies and other factors that lead to inflation of accuracy rates, and (b) the pervasive tendency towards overconfidence in the mental health field (which is far from alone in this respect). As the title of my commentary attempts to convey, we are advancing, but are not yet advanced, in evaluating effort, and we still face multiple vexing problems that will require large-scale research efforts. Unfortunately, if or when we are more certain than is justified, or act that way, there is considerable potential for harm.

At this point in our scientific efforts, there is often minimal value in developing and investigating new tests and measures that are highly redundant with methods that are already available, or studying groups of individuals we likely already know how to classify with relative surety. If we continue to focus mainly on extreme, or relatively easy or unambiguous cases, and less so on more difficult cases, progress will be impeded. Understandably, too much ambiguity in group assignment can seriously hinder efforts to evaluate measures. However, given the current state of knowledge, if we go too far in seeking unambiguous group assignment, it will enfeeble work on cases that should be a major focus: cases we presently have trouble classifying. As one might discern by the frequency of exclusions (especially TCTC cases) in studies, these more challenging presentations make up a sizeable proportion of examinees and are the very individuals, especially in the courtroom, with whom we need the most help.

Reducing error in criterion group assignment is desirable, but only to a point, and not at the cost of: (a) excluding the cases we most need to study and (b) running study after study in which criterion groups will likely, or almost certainly, differ qualitatively and quantitatively from the difficult cases. When such studies are applied to more challenging cases in legal settings, they are likely to generate problematic error rates. If individuals are included in criterion groups because they can be identified, how can they give us guidance on the cases we do not know how to identify (or for individuals excluded from the criterion groups)? For example, if individuals in a study are assigned to a poor effort group because, among other things, they fell well below chance on a forced-choice measure, then they will likely differ from individuals excluded from the study because it was too hard to determine their proper group assignment.

I do not agree that we need gold standards to proceed or make substantial progress, although, as Jewsbury (2023) stated in his commentary, they would be a welcome development. However, such criteria are not realistic to expect at present (and probably well into the future). Rather, we must make do with valid but fallible criteria, trying to take the level of fallibility into account and adjusting for it. Using fallible criteria is commonly our lot in scientific fields or areas for varying periods of time. For example, an early measurement of temperature was based on touch, and suppose someone argued that advancement was blocked because a gold standard was lacking?

As knowledge builds through programs of research, or bootstrapping operations, the methods we originally tested against fallible indicators may be gradually refined and established as more accurate than the fallible indicators with which we started, e.g., the thermometer versus touch; intellectual testing versus teachers' impressions; EEG recordings to identify paroxysmal brain activity rather than various cruder methods; and the hundreds of other examples one can cite across the history of science. Much of this progress comes down to construct validation and the development of more rigorous models and theories that show increasing levels of verisimilitude, or are better approximations of truth.

Finally, no matter how good our methods become, practitioners and experts still need to implement them properly. Faust et al. (2021) provide over 20 examples of problem practices in effort assessment that can lead to error. For example, research shows that mental health professionals often resist the use of properly developed statistical prediction methods, despite the large volume of studies demonstrating that they almost always equal or exceed clinical judgment and thus are superior overall (*A*Egisdottir et al., 2006; Dawes et al., 1989; Grove et al., 2000). Even when well-validated statistical methods are considered, clinicians often countervail or reject the outcomes freely and often, generally leading to no better, and often diminished or markedly diminished, accuracy (Dawes et al., 1989; Faust et al., in press; Guay & Parent, 2018; Krauss, 2004; Schmidt et al., 2016; Wormith et al., 2012).

Practices for selecting and combining tests and assessment methods for appraising effort, and for interpreting results, may not comport with research findings or validated approaches. There is often minimal regulation of psychologists' assessment procedures and interpretive practices, especially when they reference clinical judgment as a guiding consideration. Practice standards and ethical guidelines often minimize responsibility for departures from the best validated methods by providing escape clauses that demand little other than the rationale that one is exercising professional judgment.

Surely, there can be compelling reasons to break from usual procedures or identify potential exceptions or uncertainties about well-supported interpretive methods. However, when countervailing the best-validated procedures becomes the modus operandi or final arbiter almost no matter the scientific findings, the practitioner is very likely underutilizing or undermining the best knowledge in our field. One might think that experts who routinely break from science and instead depend heavily on unverified procedures would face professional opposition or sanction, but this outcome seems to be infrequent. (Excluding experts from testifying, which is too detailed a matter to get into here, is also relatively rare, especially in state court.) It might be helpful if the authors of test manuals, who may readily give experts an easy out by describing professional judgment as the decisive factor in decision making, at least counterbalanced such affirmations with resolute statements about the potential, if not likely, negative outcomes of countervailing well-supported methods routinely.

Some Areas of Potential Disagreement or Concern

Some areas of possible disagreement or concern I have with Leonhard's work have already been discussed, and in this section, I wish to raise a few additional issues or expound on a few earlier points. To begin, in his second article, Leonhard states that "Computations presented in the present companion reviews should therefore be considered for their qualitative probative value regarding the statistical and research methods examined, rather than as quantitative estimates of actual classification accuracy (p. XX)." I find this statement somewhat confusing because the two articles provide substantial quantitative information. Are these figures to be viewed at least as approximations, and if not, how are we to interpret and use them? How much, or what, probative value might they have? I am not questioning the value of illustrating key methodological problems. Such illustrations can help counter overconfidence and expert testimony overstating the trustworthiness of methods and research outcomes, and can serve to pinpoint critical research issues and reshape efforts productively. It might be helpful if Leonhard further explains his cautionary statement and what he believes we can and cannot take from his articles and analyses.

Leonhard's cautions about probative value as opposed to taking quantitative results he generated at face value are particularly well advised in his analysis of TCTC cases and the impact of their exclusion from studies (see in particular Table 2 of Part II, p. XX). I do not take issue with the more general and important points Leonhard raises about the high rates of exclusion in multiple studies and their marked impact in inflating accuracy rates. These are critical points that demand our attention and lead to serious problems, such as experts drawing and expressing far more confident conclusions than are warranted, or just plainly getting things wrong. However, Leonhard's analysis assumes that all eliminated cases would have been misclassified, which could only occur if we had perfect methods and always selected the opposing outcome.

It might have been helpful if Leonhard calculated accuracy rates in comparison to chance, but that was not the case. Rather, treating every TCTC case as if it were misclassified overestimates potential impact (an impact that is problematic enough as it is). Treating all TCTC cases as errors can also lead to misrepresentative results when ordering the relative efficacy of methods because rates of exclusion, which differ across studies, will appear to have greater impact in reducing accuracy than is actually the case. In addition, if researchers implement stringent criteria for group assignment, then a certain percentage of cases may not meet these requirements yet still be classified correctly in most cases or at least at levels greater than chance. Additional or alternative procedures for measuring the influence of case exclusion could add important perspective to the overall and comparative magnitude of this critical problem across studies and methods, and extend Leonhard's already valuable contributions.

Leonhard's proposals for determining prior odds seem problematical, although this is surely a challenging problem, and Jewsbury's commentary (2023) provides more in-depth coverage and insight than I can offer. Leonhard, as already noted, acknowledges limits in dichotomous classification, which is a point worth re-emphasizing given the resultant oversimplification of various complicated problems and the need to apply alternative statistical methods to appraise how well tests and assessment approaches perform. For example, the value of determining accuracy rates for dichotomous classifications can be restricted by a number of other factors and considerations, such as the frequency of mixed presentations (e.g., injured and malingering). It would also be useful to increase our proficiency in measuring such dimensions and phenomena as where level of effort falls on a continuous scale (which will likely require use of measures with higher ceilings than a test like the TOMM); the extent to which performance on effort tests can predict the magnitude and presence of reductions in performance on other standard tests or in specific areas; whether we can develop ways to adjust scores on at least some standard tests based on effort test performance, especially if reduced effort is not too extreme or global; whether or in what forms low effort might have taxometric status; and how to better distinguish intercorrelations reflecting mere redundancy or compounded error as opposed to incremental validity.

I wish Leonhard, perhaps in his response to commentary, might place greater emphasis on the distinction between compromised effort and malingering. His choice of terminology for his articles is understandable, but it risks strengthening questionable assumptions about what the types of measures at issue actually assess or can differentiate. Given the array of factors that may reduce effort, the difficulties that are often present in distinguishing among various potential causes, and the risks created for false-positive and false-negative identifications of malingering, the continued practice of some professionals to refer to these measures as malingering tests probably should be modified. The gradual shift, and now broadly recommended practice, of describing these tests and assessment methods as performance validity measures is a welcome step, although it might not go quite far enough.

Courts might sometimes take a closer look at the ways validity tests and methods are described. Lawyers might also wish to challenge the admissibility of referring to such methods as malingering tests or the use of similar terms, especially given the current state of knowledge and the limitations that are often present in distinguishing malingering from reduced effort. Certainly, at times, these distinctions are fairly obvious or clear, such as when a person performs well below chance on a number of measures and abysmally low on various other effort tests, yet is holding a high level job and handling life challenges with minimal difficulty. However, in many cases, parsing the cause or causes that account for low effort or standing on possible qualitative indicators (e.g., inaccurate presentation of history, false attributions) are beyond our current scientific knowledge and assessment methodologies.

Leonhard raises important questions about proposals for diagnostic criteria and the Sweet et al. (2021) criteria in particular. Additional emphasis might be placed on the introduction of subjective elements in such proposals that create further concerns and can compound other potential problems. Subjective elements might include appraising the presence of marked discrepancies between test data and symptom reports and other evidence and the inability to account for behaviors meeting criteria for invalid presentation by another developmental, medical, or psychiatric condition. Depending on specifics, such criteria can increase error, especially to the extent they can be difficult to evaluate or are less valid than other criteria. Neuropsychological test results, for example, except when extreme, often predict everyday functioning with only modest validity. Consequently, finding discrepancies between the two is not unusual, and an expert who conducts a thorough search for inconsistencies will almost always find some.

The criterion addressing the inability to account for indicators of invalid presentation by other causes, such as medical or psychiatric conditions, can create other difficulties. There are surely cases, and perhaps even relatively frequent ones, in which cause is difficult to determine. Using default criteria when scientific knowledge of causal elements is limited for a range of presentations can easily lead to error. One might consider, for example, how often individuals with neurological diseases were diagnosed with psychological conditions until knowledge improved. In addition, these rule outs can take on circular elements because a litigant might withhold information about psychological or medical problems to direct attention and causal attribution to brain impairment.

Various potential “red flags” for evaluating effort may have solid conceptual foundations, or there may be good reasons to think they will work, but they are often under-researched. Sound methods for evaluating their presence or degree and determining their value are often lacking, especially when compared to other, and potentially stronger, variables. Incremental validity depends on both the accuracy of methods and whether they add to, or how well they compare to, other predictors. Hence, even when valid variables are added to other predictive variables, if the former are weaker than the latter, they may not increase overall accuracy and can even lower it. Often, knowing what to include and what to exclude, and in the latter case, what may be valid but weaker variables, maximizes accuracy, whereas including weaker variables can extract a considerable cost.

Perhaps where I would take strongest issue is with Leonhard’s description of construct validity, and in particular, appearing to reduce it largely to the study of correlations. It also seems, although perhaps I am mistaken, that the role of construct validity has been underemphasized or undervalued in research on effort tests. To begin with, malingering is clearly a hypothetical construct, and in my opinion will never be captured appropriately by some type of operational definition. Furthermore, if one references Cronbach and Meehl’s (1955) classic article on construct validity, they are describing something far broader than merely looking at patterns of correlations, but rather a program for investigating and developing proposed constructs and the network of assumptions, postulates, and connections that characterize them (see also Smith, 2005; Strauss & Smith, 2009). Thus, for example, experimental studies predicting outcomes based on the theoretical structure and assumptions encompassing the construct are among the ways to examine its scientific standing or verisimilitude.

Many of the methodological and conceptual problems Leonhard describes are interrelated to ambiguities in research criteria and issues that interface with construct validity. For example, issues about criteria used to assign individuals to groups in studies and the search for more definitive criteria are related to construct validation concerns. How can one, for example, try to develop trustworthy criteria for evaluating the accuracy of classification if one lacks decent understanding of the construct under consideration? Is malingering a taxon, can it be limited to standing on dimensions, how independent might these dimensions be from one another, how many dimensions might be needed to capture key phenomena, and what qualitative or quantitative factors help in differentiating malingering from other causes of questionable or poor effort? In classifying forms of validity, construct validity is often placed as the super-ordinate category under which other forms of validity fall. One could argue that such a scheme is an overreach, which I believe to be the case, and that in various situations, our main interest is simply predictive or criterion-based accuracy. However, when we are dealing with hypothetical constructs rather than events or physical entities, as is the case with malingering and effort, construct validity is often a central concern and a critical component of scientific efforts.

Acknowledgements I wish to express my thanks to Kathleen Monahan for her invaluable help in the preparation of this manuscript.

Author Contribution Not applicable.

Availability of Data and Materials Not applicable.

Declarations

Ethical Approval Not applicable.

Competing Interests The authors declare no competing interests.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Bhowmick, C., Hirst, R., & Green, P. (2021). Comparison of the Word Memory Test and the Test of Memory Malingering in detecting invalid performance in neuropsychological testing. *Applied Neuropsychology: Adult, 28*(4), 486–496. https://doi.org/10.1080/23279095.2019.1658585

Clark, A. L., Amick, M. M., Fortier, C., Milberg, W. P., & McGlinchey, R. E. (2014). Poor performance validity predicts clinical characteristics and cognitive test performance of OEF/OIF/OND veterans in a research setting. *The Clinical Neuropsychologist, 28*(5), 802–825. https://doi.org/10.1080/13854046.2014.904928

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgement. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Faust, D., & Ahern, D. C. (2012). Clinical judgment and prediction. In D. Faust (Ed.), *Coping with psychiatric and psychological testimony* (6th ed.) (pp. 147–208). Oxford University Press. https://doi.org/10.1093/med:psych/9780195174113.003.0009

Faust, D., Arkes, H. R., & Gaudet, C. E. (in press). Applying decision research to improve clinical outcomes, psychological assessment, and clinical prediction. Oxford University Press.

Faust, D., & Furman, A. (2022). When clinical judgment and science conflict, how does one decide? The epistemological status of learning from experience vs. science. In C.L. Cobb, S.J. Lynn, & W. O'Donohue (Eds.), *Toward a Science of Clinical Psychology: A Tribute to the Life and Works of Scott O. Lilienfeld* (pp. 71–104). Springer, Cham. https://doi.org/10.1007/978-3-031-14332-8_5

Faust, D., Gaudet, C. E., Ahern, D. C., & Bridges, A. J. (2021). Assessment of malingering and falsification: Continuing to push the boundaries of knowledge in research and clinical practice. In A.M. Horton & C.R. Reynolds (Eds.), *Detection of malingering during head injury litigation* (Volume 1, pp. 1–156.) Springer, Cham. https://doi.org/10.1007/978-3-030-54656-4_1

Faust, D., Hart, K., & Guilmette, T. J. (1988). Pediatric malingering: The capacity of children to fake believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology, 56*(4), 578. https://doi.org/10.1037//0022-006x.56.4.578

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Guay, J. P., & Parent, G. (2018). Broken legs, clinical overrides, and recidivism risk: An analysis of decisions to adjust risk levels with the LS/CMI. *Criminal Justice and Behavior, 45*(1), 82–100. https://doi.org/10.1177/0093854817719482

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2004). Adapting educational and psychological tests for cross-cultural assessment. *Psychology Press*. https://doi.org/10.4324/9781410611758

Heaton, R. K., Ryan, L., & Grant, I. (2009). Demographic influences and use of demographically corrected norms in neuropsychological assessment. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric and neuromedical disorders* (pp. 127–155). Oxford University Press.

Jewsbury, P. A. (2023). Invited commentary: Bayesian inference with multiple tests.

Jewsbury, P. A., & Bowden, S. C. (2014). A description of mixed group validation. *Assessment, 21*(2), 170–180. https://doi.org/10.1177/1073191112473176

Kovach, S. (2017). The effect of coaching on the ability to identify and pass a measure of insufficient effort. (Unpublished master's thesis). University of Rhode Island.

Krauss, D. A. (2004). Adjusting risk of recidivism: Do judicial departures worsen or improve recidivism prediction under the Federal Sentencing Guidelines? *Behavioral Sciences & the Law, 22*(6), 731–750. https://doi.org/10.1002/bsl.609

Leonhard, C. (2023). Review of statistical and methodological issues in the forensic prediction of malingering from validity tests: Part I: Statistical issues. *Neuropsychology Review.*

Miller, D. J., Spengler, E. S., & Spengler, P. M. (2015). A meta-analysis of confidence and judgment accuracy in clinical decision making. *Journal of Counseling Psychology, 62*(4), 553–567. https://doi.org/10.1037/cou0000105

Mulenga, K., Ahonen, T., & Aro, M. (2001). Performance of Zambian children on the NEPSY: A pilot study. *Developmental Neuropsychology, 20*(1), 375–383. https://doi.org/10.1207/S15326942DN2001_4

Schmidt, F., Sinclair, S. M., & Thomasdóttir, S. (2016). Predictive validity of the youth level of service/case management inventory with youth who have committed sexual and non-sexual offenses: The utility of professional override. *Criminal Justice and Behavior, 43*(3), 413–430. https://doi.org/10.1177/0093854815603389

Sherry, N., Ernst, N., French, J. E., Eagle, S., Collins, M., & Kontos, A. (2022). Performance validity testing in patients presenting to a specialty clinic with a mild traumatic brain injury. *Journal of Head Trauma Rehabilitation, 37*(3), E135–E143. https://doi.org/10.1097/HTR.0000000000000692

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making, 18*(1), 29–53. https://doi.org/10.1002/bdm.486

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17*(4), 396–408. https://doi.org/10.1037/1040-3590.17.4.396

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., ... & Conference Participants. (2021). American Academy of Clinical Neuropsychology (AACN) 2021 consensus statement on validity assessment: Update of the 2009 AACN consensus conference statement on neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, *35*(6), 1053–1106. https://doi.org/10.1080/13854046.2021.1896036

Tombaugh, T. N. (1996). *Test of memory malingering*. Multi-Health Systems.

Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M. J. (2012). An investigation of self-assessment bias in mental health providers. *Psychological Reports, 110*(2), 639–644. https://doi.org/10.2466/02.07.17.PR0.110.2.639-644

Weschler, D. (2008). WAIS-IV technical and interpretive manual. Pearson.

Wormith, J. S., Hogg, S., & Guzzo, L. (2012). The predictive validity of a general risk/needs assessment inventory on sexual offender recidivism and an exploration of the professional override. *Criminal Justice and Behavior, 39*(12), 1511–1538. https://doi.org/10.1177/0093854812455741

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.