Wolfe, P.L., Millis, S.R., Hanks, R., Fichtenberg, N., Larrabee, G.J., & Sweet, J.J. (2010). Effort indicators within the California Verbal Learning Test-II (CVLT-II). *The Clinical Neuropsychologist*, *24*, 153–168. doi:10.1080/13854040903107791

Wu, T.C., Allen, M.D., Goodrich-Hunsaker, N.J., Hopkins, R.O., & Bigler, E.D. (2010). Functional neuroimaging of symptom validity testing in traumatic brain injury. *Psychological Injury & Law*, *3*, 50–62. doi:DOI 10.1007/s12207-010-9067-y

doi:10.1017/S1355617712000409

DIALOGUE RESPONSE

Response to Larrabee

Erin D. Bigler

Neuropsychology needs objective methods that confidently and accurately reflect the validity of brain-behavior relationships as measured by neuropsychological assessment techniques. Symptom validity testing (SVT) has emerged as a method designed to address validity of neuropsychological test performance; but, just like the field of neuropsychology, SVT research is new and evolving. Within any new research endeavor, first generation studies often demonstrate broad support for a new construct but as the research expands more complex issues arise that require refinements in theory and practice (Oner & Dhert, 2011). Such is the case with SVT research and its clinical application. One goal of the dialogue with Larrabee on the current status of SVT research and clinical application was to highlight areas of agreement and disagreement. My review challenges some SVT assumptions, pointing out the need for refinements in methods and theory, calling for improved research designs that will hopefully lead to a more complete understanding of SVT use and interpretation in neuropsychological assessment.

Larrabee (this issue), in response to my SVT review (see Bigler, this issue), argues for a change in terminology, abandoning the singular term "effort" in favor of "performance validity" and "symptom validity" and offers cogent reasoning and research to support such a distinction. In my opinion, the term effort as a singular descriptor in neuropsychology should be abandoned in favor of the performance validity and symptom validity terms as suggested by Larrabee in his commentary. As already stated in the critique there are simply too many potential meanings suggested with just the term effort or "effort tests," spanning the biological to inferring intent. From the biological, effort suggests neural factors associated with basic drives and emotional states (see Sarter, Gehring, & Kozak, 2006). Within cognitive neuroscience, effort relates directly to complexity of stimulus processing (Kohl, Wylie, Genova, Hillary, & Deluca, 2009) and levels of motivation (Bonnefond, Doignon-Camus, Hoeft, & Dufour, 2011; Harsay et al., 2011). In forensic and applied neuropsychology, the effort term suggests some intention on the subject's part where poor effort may be equated with malingering (see Williams, 2011). These multiple meanings make the term imprecise when used in neuropsychological parlance to describe test behavior. The "performance validity" and "symptom validity" terminology represent far more accurate descriptors of what is being assessed and neuropsychology will be better served by following Larrabee's recommendation.

There are also two basic agreements on what may be considered SVT tenets: (1) questions of "symptom" and "performance" invalidity are proportional to the number SVT items ***not*** passed and, (2) close to or below chance SVT test performance levels are the clearest and most indisputable indicators for invalidity. In my opinion, little debate about the above two points is needed. For forced-choice SVT measures, invalid neuropsychological test performance may be assumed as SVT performance falls substantially below a conventionally established cut-score. SVT performance at, near, or below chance reflects invalid test performance.

Despite these points of agreement, two major SVT topics where our opinions diverge are: (1) the "false positive/false negative problem and interpretative validity issues" and, (2) the "rigor" of SVT study designs.

THE FALSE POSITIVE PROBLEM AND INTERPRETATIVE VALIDITY ISSUES

The most effective SVT will minimize false positive and negative classifications with the false positive typically being the more serious error. False positive classification occurs when failed SVT scores are used to designate invalid neuropsychological test performance when in fact, the "failed" SVT performance occurs *because* of the underlying neurological and/or neuropsychiatric condition. The clinical gravity of a false positive SVT decision for neuropsychology is obvious—in the face of a false positive SVT indicating

invalidity of neuropsychological test findings, proper clinical diagnosis, service, and treatment may be improperly made, withheld, denied, or delayed.

As a profession neuropsychology needs to make sure that the best research informs the clinician and/or researcher with the most complete and correct information for making SVT interpretive statements. As pointed out in the critique, several SVT failures in the Locke, Smigielski, Powell, and Stevens (2008) study—all participants of whom were not in litigation and had been independently diagnosed with an acquired brain injury—performed within the "near miss" zone of SVT performance. Do these SVT scores truly reflect invalid performance across *all* neuropsychological tests administered that cannot be explained by their neurological/neuropsychiatric condition? Is there something unique about their injuries that lowered their performance on the SVT? How many of these subject's SVT scores represent a true false positive and how would the neuropsychologist know?

These are important questions without answers. The very nature of an SVT cut-score is to make a dichotomous decision and if that SVT cut-score is applied to all types of neurological/neuropsychiatric conditions this becomes a "one-size-fits-all" approach. The Diagnostic and Statistical Manual–IV edition (DSM-IV) lists 17 Axis I or II general categories with over 450 separate diagnostic codes and about that many International Classification of Diseases, Tenth Revision (ICD-10) classifications involving neurological disorders are also listed in the DSM-IV (see American Psychiatric Association, 1994). Unanswered SVT questions thereby remain as to whether SVT findings broadly apply across all DSM-IV classifications; whether different SVTs should be used depending on the disorder being assessed; when in the assessment protocol should an SVT be administered (first test administered, somewhere in the middle, multiple ones, does not matter, etc.?); whether different cut-scores apply for different patient demographics, etc.? More research is needed to address these basic SVT questions and others not listed.

THE QUALITY OF SVT RESEARCH DESIGN

Larrabee spends a good deal of his commentary defending the rigor of SVT research. As already stated there is sufficient convergence and quality of research to support the two broad SVT tenets stated above. Issues of research design quality are not directed at these fundamental points.

The opening statements of the American Academy of Clinical Neuropsychology (AACN) document on effort, response bias, and malingering discuss the necessity of ever improving research designs to advance the field where Heilbronner, Sweet, Morgan, Larrabee, and Millis (2009) state the following, " … science-driven healthcare specialties create progress by a process of challenging current and new ideas through intellectual discourse and empirical hypothesis testing" (p. 1094). McGrath and colleagues (see McGrath, Kim, & Hough, 2011; McGrath, Mitchell, Kim, & Hough, 2010), in their reviews and commentaries of response bias research within applied psychological assessment, emphasize that SVT research designs must be the most "stringent" before any blanket acceptance of SVT interpretive statements can be made, especially in terms of Type II statistical errors. Given these guidelines, it seems a non-arguable point that neuropsychology seeks the best designed, most rigorous studies from which to base applied decision making. The better the research design the more generalizable are the findings.

As pointed out in Bigler (this issue), research design rigor straightforwardly can be assessed using the American Academy of Neurology (AAN) rating method (see Edlund, Gronseth, So, & Franklin, 2004). As a historical note, it was this method of rating quality of neuropsychological research involving cases of dementia, cerebrovascular disease, traumatic brain injury (TBI) and epilepsy, that in 1996 allowed the AAN Therapeutics and Technology Assessment (TTA) subcommittee (see American Academy of Neurology, 1996) to grant a Class II, "Type A" rating for using neuropsychological assessment techniques to evaluate the cognitive and neurobehavioral effects of these specific neurological conditions. A Type A rating means that the technique is "established as useful/predictive for a given condition in the specified population." (p. 598). The AAN publication predates the development of current SVT methods although some response bias and validity issues were discussed in the AAN statement. AAN guidelines are clear that Type A ratings come *only* after "established Class I or II designed studies" and then only after a comprehensive review by the TTA subcommittee. By AAN research design classification standards, Class I is the most rigorous with Class IV the least.

Cappa, Conger, and Conger (2011) provide guidelines for another method of rating experimental design quality for neuropsychological outcome research by assessing nine points related to study design. These nine points are summed to create four classifications from best to worst as follows: commendable, adequate, marginal and flawed.

Regardless of whether the AAN (1996) guidelines or those from Cappa et al. (2011) are used to rate rigor of SVT study design, the best designed studies ("Class I" or "commendable") will be those with *a priori* defined criteria that require a prospective experimental design, uniform recruitment where investigators and/or clinicians have well defined and independent roles especially in diagnostic decision making and classification—along with the study being appropriately blinded including all aspects of data coding, entry and analysis—to list just some of the key elements. By AAN standards, Class II may include retrospective studies but still requires investigator independence, blinded assessments, and data analysis. Class III may be retrospective and partially unblinded but still requires independence of the investigators. Class IV may include case series and be based on expert opinion where non-independence of the investigators is present.

Most of the SVT research cited in my review and Larrabee's comment would merit no better than a Class III level AAN rating or an adequate-to-marginal rating by Cappa et al. standards. As pointed out by Edlund et al. (2004), Class III

https://doi.org/10.1017/S1355617712000409 Published online by Cambridge University Press

and IV level research is important for hypothesis building and proof of concept studies. Clearly, solid SVT research has been done; that is why both the AACN (Heilbronner et al., 2009) and the National Academy of Neuropsychology (NAN) (see Bush et al., 2005) have position papers on the use of SVT measures. This does not mean as a profession we should be content with Class III and IV level SVT research and as clinicians and researchers not demand better designed studies.

Larrabee makes the point about the importance of "known group" or "criterion" design as an example of the rigor of existing SVT investigations. However, as pointed out in the AACN consensus statement, specifically in the section on known groups, Heilbronner et al. (2009) point out "…. Developing appropriate external criteria for defining response bias can be a major methodological challenge" (p. 1118). For example, several SVT studies have used a forensic sample establishing a known group with "objective" brain damage demonstrated by radiological evidence of abnormality. However, careful reading of these studies show that the determination of who is in the "objectively brain damaged group" is based entirely on the retrospectively obtained clinical record and whatever radiological report the author/investigator may have available. None of these studies provide any quality control over the neuroimaging method used, the sensitivity of the neuroimaging tool to detect the problem, or the radiologist making the rating. So without the uniformity that comes from exactly the same procedure prospectively performed on all subjects, these known groups with "objective indicators of brain damage" versus "no objective indicators of brain damage" become ill-defined and potentially meaningless. Retrospective data sets based on forensic or clinical samples will never be Class I or II or "commendable" research designs. Better SVT research, prospectively designed, and independently conducted is needed.

In the spirit of the 2009 AACN recommendations on the assessment of effort, response bias, and malingering "….progress …" is made via "…. a process of challenging current and new ideas through intellectual discourse and empirical hypothesis testing" (Heilbronner et al., 2009, p. 1094). This is the challenge for the next generation of SVT studies—better research design, less reliance on samples of convenience, and a focus on prospectively designed, independently conducted investigations.

REFERENCES

American Academy of Neurology. (1996). Assessment: Neuropsychological testing of adults. Considerations for neurologists. Report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. *Neurology*, *47*, 592–599.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders–4th edition (DSM-IV)*. Washington, DC: American Psychiatric Association.

Bigler, E.D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*, 632–642.

Bonnefond, A., Doignon-Camus, N., Hoeft, A., & Dufour, A. (2011). Impact of motivation on cognitive control in the context of vigilance lowering: An ERP study. *Brain and Cognition*, *77*, 464–471. doi:10.1016/j.bandc.2011.08.010

Bush, S.S., Ruff, R.M., Troster, A.I., Barth, J.T., Koffler, S.P., Pliskin, N.H., … Silver, C.H. (2005). Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology*, *20*, 419–426. doi:10.1016/j.acn.2005.02.002

Cappa, K.A., Conger, J.C., & Conger, A.J. (2011). Injury severity and outcome: A meta-analysis of prospective studies on TBI outcome. *Health Psychology*, *30*, 542–560. doi:10.1037/a0025220

Edlund, W., Gronseth, G., So, Y., & Franklin, G. (2004). *Clinical practice guideline process manual: For the Quality Standards Subcommittee (QSS) and the Therapeutics and Technology Assessment Subcommittee (TTA)*. St. Paul: American Academy of Neurology.

Harsay, H.A., Cohen, M.X., Oosterhof, N.N., Forstmann, B.U., Mars, R.B., & Ridderinkhof, K.R. (2011). Functional connectivity of the striatum links motivation to action control in humans. *The Journal of Neuroscience*, *31*, 10701–10711. doi:10.1523/jneurosci.5415-10.2011

Heilbronner, R.L., Sweet, J.J., Morgan, J.E., Larrabee, G.J., & Millis, S.R. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, *23*, 1093–1129. doi:10.1080/13854040903155063

Kohl, A.D., Wylie, G.R., Genova, H.M., Hillary, F.G., & Deluca, J. (2009). The neural correlates of cognitive fatigue in traumatic brain injury using functional MRI. *Brain Injury*, *23*, 420–432. doi:10.1080/02699050902788519

Locke, D.E., Smigielski, J.S., Powell, M.R., & Stevens, S.R. (2008). Effort issues in post-acute outpatient acquired brain injury rehabilitation seekers. *Neurorehabilitation*, *23*, 273–281.

McGrath, R.E., Kim, B.H., & Hough, L. (2011). Our main conclusion stands: Reply to Rohling et al. (2011). *Psychological Bulletin*, *137*, 713–715. doi:10.1037/a0023645

McGrath, R.E., Mitchell, M., Kim, B.H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*, 450–470. doi:10.1037/a0019216

Oner, F.C., & Dhert, W.J. (2011). Challenging the medico-industrial-administrative complex. *The Spine Journal*, *11*, 698–699. doi:10.1016/j.spinee.2011.08.010

Sarter, M., Gehring, W.J., & Kozak, R. (2006). More attention must be paid: The neurobiology of attentional effort. *Brain Research Reviews*, *51*, 145–160. doi:10.1016/j.brainresrev.2005.11.002

Williams, J.M. (2011). The malingering factor. *Archives of Clinical Neuropsychology*, *26*, 280–285. doi:10.1093/arclin/acr009