INTRODUCTION

A large number of scoring systems are developed for psychiatry. The Global Assessment of Functioning (GAF) is known worldwide, translated into many languages, and used in many outcome studies (Aas, 2010; 2011; 2014). GAF is used to rate severity of illness in psychiatry and covers the range from positive mental health to severe psychopathology. It is an overall (global) measure of how patients are doing (Moos et al., 2000; Rosse & Deutsch, 2000). GAF is not LQWHQGHG WR EH D GLDJQRVLV VSHFL¿F VFRULQJ V\VWHP EXW D JHQHUL F Compared to diagnosis, GAF values represent more multidimensional information (Rosse & Deutsch, 2000; Schorre & Vandvik, 2004). The degree of mental illness is measured by rating psychological, social and occupational functioning (Goldman et al., 1992; Vatnaland et al., 2007). The simplicity of GAF is an advantage (Aas, 2010).

Internationally, recording GAF is either done with a single value (this is the most severe of the symptom and functioning values) or both symptom (GAF-S) and functioning (GAF-F) values are recorded. The symptom and functioning scales have both 100 scoring possibilities (1-100). The 100-point scales are divided into 10 intervals, or sections, each with 10 scoring possibilities (examples: 31-40 and 51-60). Verbal instructions (called anchor points) describe symptoms and functioning relevant for scoring in the 10-point intervals. The anchor points represent hierarchies of mental illness (McDowell & Newell, 1987; Pedersen et al., 2007; Vatnaland et al., 2007). The anchor points for interval 1-10 describe the most severely ill and the anchor points for interval 91-100 describe the healthiest. In addition to anchor points, examples are found for each 10-point interval. The examples are intended to help with the scoring in each interval. For example, in the interval 51-60 (moderate symptoms) on the symptom scale, patients with occasional panic attacks can be rated, and in the LQWHUYDO PRGHUDWH GLI¿FXOW\ L IXQFWLRQLQJ RQ WKH IXQFWLRQLQJ VF peers or co-workers and few friends can be rated (Karterud et al., 6FKRUUH 9DQGYLN 7KH ¿QHU (for example 32, 35, 37 and 55, 57, 59) provides the possibility of distinguishing between nuances (Thomson, 1989), but there are no verbal instructions for this grading found on either the two scales. Research on GAF shows problems with both reliability and validity. Reliability studies show the extreme 20% of raters to account for more than 50% of the spread of scores and deviations can be 20 points or more (Loevdahl & Friis, 1996; Vatnaland et al., 2007). Different studies show inter-rater reliability to be highly variable, but it should be noticed that this includes very good reliability. Reliability seems to be lower in routine clinical practice than in research (Burlingame et al., 2005; Hilsenroth et al., 2000; Moos et al., 2000; Soderberg et al., 2005; Startup et al., 2002; Vatnaland et al., 2007). Concurrent validity (Bates et al., 2002; Burlingame et al., 2005; Goldman et al., 1992; Hall, 1995; Hay et al., 2003; Hilsenroth et al., 2000; Jones et al., 1995; Niv et al., 2007; Patterson & Lee, 1995; Pedersen et al., 2007; Piersma & Boes , 1997; Robert et al., 1991; Roy-Byrne et al., 1996; Salvi et al. 2005; Tungstrom et al., 2005) and predictive

1992; Hay et al., 2003; Moos et al., 2000; Niv et al., 2007; Parker et al., 2002) are more problematic. There are few empirical results for GAF sensitivity (Bird et al., 1987).

In the clinic, the primary goal of the assessment process is to contribute to the solution of a person's problems (Bruyn, 2003). A generic and global scoring system, such as GAF, that covers the range from positive mental health to severe psychopathology has advantages for clinical practice (for example, routine quality assessment of treatment, supplementing scales that give more detail) (Lingjaerde et al., 1989), research (for example, comparison of treatment outcome across diagnoses), and policy and management planning (for example, allocation of resources, measurement of case-mix in psychiatric organizations). We are dealing with a wide range of potential applications and GAF must be good enough for its purposes. To dismiss an existing instrument due to problems can be a too simple solution (Streiner & Norman, 1994). Work to improve GAF is an alternative. Further development for GAF means work to improve validity and reliability, and to ensure good sensitivity, and generic properties.

e present study is based upon the rst of three systematic literature reviews (Aas, 2010; 2011; 2014). e purpose of the study is to show the gaps in current knowledge, and ideas about further development when it comes to properties of the GAF scale.

PROPERTIES OF GAF AND GAPS IN KNOWLEDGE

3URSHUWLHV RI *$) DUH GH¿QHG DV FKDUDFWHULVWLF WUDLWV RU DWWULEXWH WKDW VHUYH WR GH¿QH *$) RU PD\ KDYH D UROH WR GH¿QH D IXWXUH QHZ *$) 7KH JDSV LGHQWL¿HG LQ WKH SUHVHQW VWXG\ DUH GH¿QHG DV properties of GAF where no, or little, research has been done, with characteristics that suggest further development is likely to have a role for improvement of GAF.

7KH ¿UVW RI WKH WKUHH V\VWHPDWLF OLWHUDWXUH UHYLHZV $DV shows the properties of the GAF scale in four main categories. These main categories (including subcategories) are important when it comes to further development of GAF and further development means work to improve GAF. The four main categories are: (1) scaling; (2) the anchor points of GAF; (3) scoring within 10-point intervals; and (4) the number of scales.

Scaling

For science in general, measurement and scaling are fundamental, but not less important for evaluation of interventions in health care. 3UREOHPV ZLWK TXDQWL¿FDWLRQ SOD\ D Z*á ZLW `p“ îwWPO 0ð á€q¾Hîà¯:ª “•¾HîÙ5sô {ðk £

having extended instructions for the upper part of the scale can be raised.

Gap in knowledge

Systematic testing of different changes in the number of anchor points with examples, and their distribution over the total scale, to REWDLQ D EHWWHU *$) LV GLI¿FXOW WR ¿QG LQ WKH KLVWRU\ RI *$)

Scoring within 10-Point Intervals

Endicott et al 1976 (Endicott et al., 1976) and the manual for DSM-IV-TR give instructions for scoring within 10-point intervals, but instructions are limited (Aas, 2011).

Gap in knowledge

Systematic study to improve scoring within 10-point intervals is limited. Categorical scales could be evaluated for the purpose. Such application of categorical scaling would require consideration of the nature and number of categories.

The Number of Scales

In the DSM-IV-TR instructions, raters are told to record only RQH ¿JXUH IRU *$) EXW ERWK V\PSWRPV DQG IXQFWLRQLQJ VKRXOG EH HYDOXDWHG 7KH SUREOHP ZLWK UHFRUGLQJ RQO\ RQH ¿JXUH LV D UHGXFWLRQ ODFN RI NQRZOHGJH LI WKH ¿JXUH LV D IXQFWLRQLQJ RU V\PSWRP VFRUH

GAF with two scales

In psychiatry, symptoms and functioning are often closely related (Goldman et al., 1992; Hilsenroth et al., 2000; Moos et al., 2000; Moos et al., 2002), but have been proposed to deviate frequently enough to recommend measuring both in outcome studies (Bacon et al., 2002; Goldman et al., 1992). GAF-S and GAF-F can be correlated with r= 0.61 (Pedersen et al., 2007).

Gap in knowledge

Symptoms and functioning are different dimensions, but knowledge about the advantage using GAF-S and GAF-F separately is limited. GAF-S and GAF-F score different dimensions, but the scores should still be correlated. Search for the right combination of GH¿QLWLRQV RI *$) 6 DQG *$)) LV OLPLWHG 0RUH VWXG\ VKRXOG EH done of reliability and validity for both GAF-S and GAF-F scales individually.

FURTHER DEVELOPMENT FOR SCALE PROPERTIES

The history of GAF does not show the research-based development of GAF to be especially strong, particularly in the context of its widespread use. Little study of systematic variation in system properties has been carried out. Many alternative forms of a new GAF could be examined (with both with major and minor FKDQJHV ,W LV GLI¿FXOW WR IRUHFDVW ZKLFK FKDQJHV DUH OLNHO\ WR SURYLG WKH PRVW VLJQL¿FDQW LPSURYHPHQWV)RU ZRUN ZLWK D QHZ *$) VRPH overall goals can be formulated:

(1) GAF should continue to be an overall (i.e. global) measure of how patients are doing; (2) with a future GAF, it should be possible to rate severity from the most severe mental illness to perfect KHDOWK *$) VKRXOG FRQWLQXH QRW EH D GLDJQRVLV VSHFL¿F VFRULQJ system, but the generic properties should be improved; (4) results from GAF scoring should continue to add information compared to what diagnoses give; (5) for a new GAF, reliability should not be lower, but rather improved; (6) work with a new GAF should aim at improved validity; (7) sensitivity should be analysed, compared to other scaling methods, and found to be good enough for the SXUSRVH FOLQLFLDQV VKRXOG ¿QG D QHZ *$) WR PDNH VHQVH DQG scoring with a new GAF should be little work requiring, i.e. scoring should be fast and easy. The goals are ambitious, but not necessarily impossible to combine.

CONCLUSIONS

No doubt, GAF has a history with limited change of basic properties. It is too simple to believe that improvement work should not be done because GAF is good enough. An international research programme with study of effects of different changes in basic properties may well be important, but is lacking. Research on basic properties has not at all played an important role for further development of GAF. Problems with GAF may be related to this. Future research could improve GAF.

REFERENCES

Aas, I. H. M. (1991). Poliklinikker og dagkirurgi. Virksomhets-beskrivelse for ambulant helsetjeneste [Outpatient clinics and daysurgery. Describing the activity of ambulatory care]. Goteborg: Nordic School of Public Health, NHV-rapport 1991, 4.

Aas, I. H. M. (2010). Global Assessment of Functioning (GAF): Properties and frontier of current knowledge. Annals of General Psychiatry 9, 20.

Aas, I. H. M. (2011). Guidelines for rating Global assessment of Functioning (GAF). Annals of General Psychiatry, 10, 2.

Aas, I. H. M. (2014). Collecting Information for rating Global Assessment of Functioning (GAF): Sources of Information and Methods for Information Collection. Current PsychiatryReviews, 10, [EPub Ahead of Print].

Bacon, S. F., Collins, M. J., & Plake, E. V. (2002). Does the Global Assessment of Functioning assess functioning? Journal of Mental Health Counseling, 24, 202-212.

Bates, L. W., Lyons, J. A., & Shaw, J. B. (2002). Effects of brief training on application of the global assessment of functioning scale. Psychological Reports, 91, 999-1006.

Bech, P., Malt, U. F., Dencker, S. J., Ahlfors, U. G., Elgen, K., Lewander, T., et al. (1993). Scales for assessment of diagnosis and severity of mental disorders. Acta Psychiatrica Scandinavica, 87(Suppl 372), 3-86.

Bird, H. R., Canino, G., Rubio-Stipec, M., & Ribera, J. C. (1987). Further measures of the psychometric properties of the Children's Global Assessment Scale. Archives of General Psychiatry, 44, 821-824.

Bowling, A. (1993). Measuring health. A review of quality of life measurements scales. Buckingham: Open University Press.

Fallmyr, O., & Repal, A. (2002). Evaluering av GAF-skaring som del av Minste Basis Datasett [Evaluation of GAF-scoring as part of Minimum Basis Dataset]. Tidsskr Nor Psykologforening, 39, 1118-1119.

)HLQVWHLQ $ 5 -RVHSK\ % 5 :HOOV & . 6FLHQWL¿F and clinical problems in indexes of functional disability. Annals of Internal Medicine, 105, 413-420.

First, M. B. (1995). Mastering DSM-IV Axis V. Journal of Practical Psychiatry & Behavioural Health, 1, 258-259.

Gelder, M., Mayou, R., & Geddes, J. (2006). Psychiatry (3rd ed.). Oxford: Oxford University Press.

Goldman, H. H., Skodol, A. E., & Lave, T. R. (1992). Revising axis V for DSM-IV: a review of measures of social functioning. The American Journal of Psychiatry, 149, 1148-1156.

Gonella, J. S. (1983). Clinical criteria for disease staging. Santa Barbara: Systemetrics Inc.

+DOO 5 & : *OREDO $VVHVVPHQW RI IXQFWLRQLQJ $ PRGL¿HG scale. Psychosomatics, 36, 267-275.

Hay, P., Katsikitis, M., Begg, J., Da Costa, J., & Blumenfeld, N. (2003). A two-year follow-up study and prospective evaluation of the DSM-IV Axis V. Psychiatry Services, 54, 1028-1030.

Hilsenroth, M. J., Ackerman, S. J., Blagys, M. D., Bauman, B. D., Baity, M. R., Smith, S. R., et al. (2000). Reliability and validity of DSM-IV axis V. The American Journal of Psychiatry, 157, 1858-1863.

Jones, S. H., Thorncroft, G., Coffey, M., & Dung, G. (1995). A brief mental health outcome scale reliability and validity of the

between the Global Assessment of Functioning and other DSM Axes in routine clinical work. *Psychiatry Services*, *56*, 439-443.

Vaillant, G. E. (2003). Mental health. *e American Journal of Psychiatry*, *160*, 1373-1384.

Vatnaland, T., Vatnaland, J., Friis, S., & Opjordsmoen, S. (2007). Are GAF scores reliable in routine clinical use? *Acta Psychiatrica Scandinavica*, *115*, 326-330.

Wells, K. B., Stewart, A., Hays, R. D., Burnam, M. A., Rogers, W., Daniels, M., et al. (1989). e functioning and well-being of depressed patients. Results from the Medical Outcomes Study. *Journal of American Medical Association*, *262*, 914-919.

Widiger, T. A., & Clark L. E. (2000). Toward DSM-V and the classi cation of psychopathology. *Psychological Bull*, *126*, 946-963.

Young, F. W. (1984). Scaling. *Annual Reviews of Psychology*, *35*, 55-81.