JOURNAL OF NEUROLOGY AND
PSYCHOLOGY RESEARCH

RESEARCH ARTICLE Open Access

M-Kappa: A New Approach to Reliability in Psychiatric Diagnostic

Dr. Martin Legind von Bergen*

MD psych. Center for Research in Psychiatric Diagnostics and Treatment, NY Vestergaardsvej 5B, 3500 Vaerelose, Denmark.

Received date: October 15, 2025, **Accepted date:** October 22, 2025, **Published date:** November 06, 2025.

***Corresponding Author:** Dr. Martin Legind von Bergen, MD psych. Center for Research in Psychiatric Diagnostics and Treatment, NY Vestergaardsvej 5B, 3500 Vaerelose, Denmark.

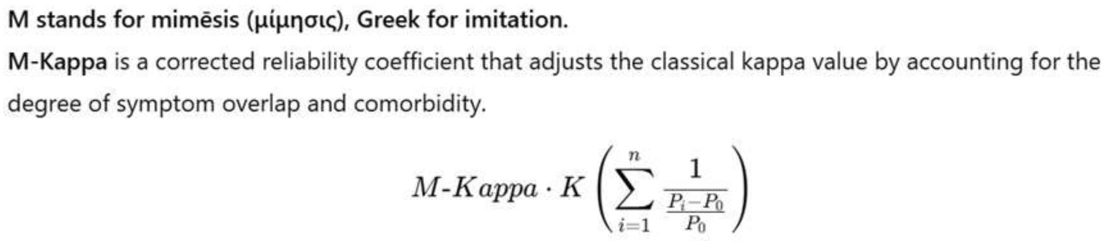
Abstract

Background: Since the introduction of the DSM-III in 1980, reliability has served as the primary metric for evaluating the utility of psychiatric diagnoses. Diagnostic categories have increasingly been defined through operational criteria, leading to greater interrater agreement as measured by Cohen's kappa (κ).

This methodological transformation also signifies a deeper epistemological shift. Reliability no longer functioned merely as a technical instrument for improving diagnostic consistency; it evolved into an epistemic norm that redefined what counted as legitimate psychiatric knowledge. In this sense, reliability supplanted validity as the discipline's guiding criterion: whereas validity concerned the ontological status and causal understanding of mental disorders, reliability became the measure of epistemic credibility. This transition marks a movement from a concern with what a diagnosis is to a concern with how consistently it can be applied — a shift from ontology to methodology.

However, many psychiatric diagnoses exhibit extensive symptom overlap and comorbidity, which undermines the scientific validity of reliability as a construct. A central issue is that reliability tends to privilege easily recognizable yet diagnostically non-specific symptoms — such as restlessness, inattention, and hyperactivity in ADHD. These symptoms occur across a wide range of psychiatric conditions as well as within normal psychological variation. Diagnoses based on such traits may yield high reliability simply due to ease of recognition, creating a false impression of precision while further eroding diagnostic validity.

In this context, reliability as measured by the kappa coefficient is merely a measure of agreement, not a Kappa, a corrected reliability coefficient, aims to restore scientific significance to kappa by accounting for variables that artificially inflate its values. M-Kappa scientifically meaningful construct. The proposed M-adjusts the classical kappa (K) to incorporate the effects of comorbidity and symptom overlap, thereby providing a more valid representation of diagnostic reliability.



M stands for mimēsis (μίμησις), Greek for imitation.

M-Kappa is a corrected reliability coefficient that adjusts the classical kappa value by accounting for the degree of symptom overlap and comorbidity.

$$M\text{-}Kappa \cdot K\left(\sum_{i=1}^{n} \frac{1}{\frac{P_i - P_0}{P_0}}\right)$$

K is the observed kappa (interrater reliability).

n is the number of comorbid conditions whose prevalence exceeds the normal base prevalence. P_0 is the baseline prevalence in the general population (typically 10% or lower). P_i is the prevalence of the ith comorbid condition in individuals with the target diagnosis.

The expression ($P_i - P_0$) / P_0 represents the relative overrepresentation of each comorbidity. The sum calculates the average relative overrepresentation across all relevant comorbidities.

When relative overrepresentation is high, the kappa is reduced proportionally, reflecting that reliability likely stems from superficial symptoms that appear across many conditions.

M-Kappa is primarily applied to the overall picture of a diagnosis. It cannot be calculated for individual symptoms but evaluating symptoms' relation to other diagnoses and the normative range can provide important insights into the diagnostic construct.

Objective: This article introduces M-Kappa – a corrected reliability measure that adjusts for non-specific symptoms and high comorbidity. M-Kappa addresses how high kappas can mask poor nosological precision.

Method: M-Kappa is calculated by adjusting kappa according to the number and degree of comorbidities whose prevalence exceeds base rates. A mathematical formula and visual model are presented to illustrate the relationship between Kappa, comorbidity, and nosological validity.

Results: The article shows that diagnoses with high kappa value, but also high comorbidity often have poor validity due to lack of diagnostic specificity. K thus represents a form of problematic reliability. The proposed inverted U-curve illustrates how rising kappas – when driven by symptom overlap – yield lower corrected kappa values via M-Kappa.

Conclusion: M-Kappa is a critical methodological tool that contributes to more accurate evaluations of reliability in psychiatric research. It allows for a nuanced understanding of when reliability is mimetic rather than substantive – and when the kappa coefficient is scientifically meaningful or misleading.

Introduction

Reliability has long been regarded as a

cornerstone of psychiatric diagnosis — a fundamental criterion for determining whether a diagnostic construct is clinically and scientifically meaningful. With the introduction of the Diagnostic and Statistical Manual of Mental Disorders, Third Edition (DSM-III) in 19801, reliability emerged as the central methodological ideal, marking a decisive shift from psychodynamic to operationalized, symptom-focused diagnostic systems.

The concept of reliability was strategically employed as a catalyst for this paradigmatic transformation. The earlier editions, DSM-I (1952)2 and DSM-II (1968)3, were grounded in psychoanalytic theory, wherein diagnoses were formulated based on unconscious motives, dream interpretation, and the clinician's subjective judgment. This approach inevitably resulted in substantial variability in diagnostic practice and, consequently, low interrater reliability.

Work on the DSM-III began in 19744, when Robert Spitzer was appointed head of the task force. However, the process was criticized for its lack of transparency: it was unclear who had selected the members and on what basis. The group was not representative of the many different schools of thought within psychiatry. After public criticism, a few psychiatrists with non-biological orientations were eventually included in the process.

Spitzer later denied that the structure of DSM-III reflected a covert agenda to promote a biological or pharmaceutical model5. Nevertheless, the system ultimately adopted a distinct medical orientation, and collaboration between the APA and the pharmaceutical industry intensified. Despite promises of an empirically grounded approach, the methodology in practice functioned largely as a consensus model—highlighting the importance of who participated in its formation.

DSM-III introduced a new structure based on symptom-defined criteria, intended to distinguish different disorders through precise inclusion and exclusion rules. Emil Kraepelin's diagnostic framework and system of classification were used as the foundational basis for the nosological structure of DSM-III.

The diagnostic process was reduced to a label derived from responses in a structured interview—a so-called "funnel model," in which complex clinical information was filtered out. The diagnosis thus became the foundation for treatment, rather than the patient's subjective account of their experiences.

Spitzer was also acutely aware of the persistent problems of reliability in psychiatric diagnosis—defined as the degree of agreement among independent clinicians. In 1960, psychologist Jacob Cohen introduced the kappa coefficient (κ) as a statistical measure of interrater reliability6 (Cohen, 1960). Unlike simple percentage agreement, κ corrects for agreement that may occur by chance; Cohen defined it as "the proportion of agreement corrected for chance"[6]. This correction was introduced because simple percentage agreement does not account for the probability that diagnosticians may arrive at the same result purely by guessing[7].

A κ value of 1.0 indicates perfect agreement between raters, 0 corresponds to agreement expected by chance, and negative values indicate systematic disagreement—meaning that independent raters tend to make opposite judgments [8-10].

Spitzer and Fleiss[4] (1974) subsequently used κ to evaluate the reliability of psychiatric diagnoses and demonstrated that levels of interrater agreement were unacceptably low, often below κ = 0.50. These findings became a central methodological justification for the development of DSM-III, in which reliability was elevated to a methodological ideal and a criterion for scientific legitimacy.

The introduction of DSM-III was therefore presented as a methodological necessity. Its primary aim was to ensure diagnostic consistency and replicability by introducing clearly defined, operational criteria.

Diagnoses were deliberately atheoretical and based solely on observable symptoms rather than etiological explanations or theoretical assumptions.

The driving force behind this development was Robert Spitzer, who, together with colleagues from Washington University, had previously developed the Research Diagnostic Criteria (RDC). The RDC introduced operational definitions, particularly for affective and psychotic disorders11 (Spitzer et al., 1975). DSM-III also built upon the Feighner Criteria12, which emphasized empirical validation and longitudinal course. Structurally, DSM-III closely reflected Kraepelinian principles of classification, conceptualizing diagnoses as discrete symptom clusters designed to predict course, treatment response, and prognosis.

This transformation was widely perceived as a process of scientification within psychiatry; however, it also entailed a marked de-emphasis of validity. Spitzer and his colleagues invoked the notion of reliability to distance themselves from the subjectivity of psychoanalytic diagnosis. Diagnostic categories were expected to be replicable regardless of theoretical orientation, and kappa coefficients became the statistical index of this interrater agreement. Yet, this shift from validity to reliability represented a theoretical prioritization rather than an empirically grounded necessity.

Subsequent critiques have repeatedly underscored the unresolved problem of validity. Kirk and Kutchins13 (1992) argued that the American Psychiatric Association (APA) failed to develop systematic methods for evaluating the validity of its diagnostic constructs. As a result, psychiatric diagnoses were reduced to clusters of symptoms devoid of clear etiological or prognostic foundations. In connection with the development of DSM-5, Thomas Insel, then Director of the National Institute of Mental Health, famously remarked that the diagnostic categories “lack validity”[14].

In retrospect, the methodological reform initiated by DSM-III did not render psychiatry more scientific in any substantive sense. By prioritizing reliability over validity, the manual replaced theoretical reflection and etiological understanding with procedural uniformity. The emphasis on interrater agreement—measured through κ coefficients—created the appearance of objectivity while concealing the conceptual fragility of the diagnostic constructs themselves [13,15].

The resulting diagnostic framework, increasingly aligned with biological and psychopharmacological models, claimed scientific legitimacy through formal standardization rather than explanatory or empirical robustness16-17 (Mayes & Horwitz, 2005; Frances, 2013). As critics have noted, this shift amounted to a reconfiguration of psychiatry’s epistemic foundations: reliability became a methodological ritual rather than a scientific achievement. The new operational paradigm substituted statistical consensus for theoretical coherence, and interrater agreement for ontological validity.

As Insel[18] (Pickersgill M.D. 2014) later observed, psychiatric categories “lack validity,” not because of insufficient measurement, but because the underlying constructs are not grounded in identifiable mechanisms or coherent theory. Thus, far from overcoming the subjectivity of psychoanalytic diagnosis, the post-1980 diagnostic system institutionalized a subtler form of subjectivity—technocratic rather than interpretive—one that masked theoretical uncertainty beneath a veneer of methodological precision. In this sense, the DSM-III revolution represents not the scientification of psychiatry, but its methodological pseudoscientification.

Reliability and the Limitations of Kappa Values

Since the introduction of DSM-III in 1980, reliability has served as the primary criterion for assessing the usefulness of psychiatric diagnoses. A diagnosis was considered legitimate if it could be applied consistently by different clinicians, regardless of their theoretical orientation. This shift toward an operationalized and symptom-focused approach brought certain methodological advantages - including greater reproducibility and increased applicability in research contexts. However, it also introduced significant challenges: reliability became not merely a methodological tool, but the very foundation upon which diagnostic constructs were built, often at the expense of validity.

Reliability and the Kappa Coefficient

In psychiatry, reliability is typically assessed as interrater agreement - that is, the degree to which two or more independent clinicians arrive at the same diagnosis. The most widely used metric for this purpose is Cohen's kappa (κ), developed by Jacob Cohen in 1960. Kappa adjusts for the level of agreement expected to occur by chance and has therefore been regarded as a more valid indicator than simple percentage agreement.

A kappa value of 1 indicates perfect agreement between raters, whereas a value of 0 signifies agreement no greater than chance. Spitzer and Fleiss4 (1974) proposed that values above 0.70 could be considered acceptable reliability — a threshold that has since become an informal standard. During the development of DSM-III, kappa was explicitly employed to exclude diagnostic categories demonstrating low interrater consistency.

In practice, however, the high kappa values reported in controlled settings have rarely been replicated in real-world clinical contexts. In a large multisite study using the Structured Clinical Interview for DSM (SCID), Williams et al. (1992)19 reported a mean kappa of only 0.47. Field trials conducted prior to DSM-5 revealed even lower coefficients: Regier et al. (2013)20 and Freedman et al. (2013)21 found kappa values below 0.50 for common diagnoses such as depression, OCD and anxiety — in some cases as low as 0.20–0.35.

The use of Kraepelin's nosological categories need not have been problematic in itself; however, the entire diagnostic enterprise was approached incorrectly from a research perspective by emphasizing reliability over validity and by employing reliability as the primary tool in the construction of DSM-III and its subsequent revisions.

These findings underscore a fundamental problem in the construction of DSM diagnostic categories: even with increasingly structured and operationalized criteria, only moderate or low levels of interrater reliability are achieved.

This methodological transformation also reflects a deeper epistemological shift. Reliability did not merely function as a technical instrument for improving diagnostic agreement; it became an epistemic norm that redefined what counted as legitimate psychiatric knowledge. In this sense, reliability replaced validity as the discipline's guiding criterion: where validity had concerned the ontological status and causal understanding of mental disorders, reliability became the measure of epistemic credibility. This transition marks a move from a concern with what a diagnosis is to a concern with how consistently it can be applied — a shift from ontology to methodology.

Reliability Without Validity

A central problem with the use of kappa values is that they tend to privilege recognizable but diagnostically non-specific symptoms — such as restlessness, agitation, or difficulties with concentration — which occur across many psychiatric conditions as well as within normal psychological variation. Diagnoses grounded in such symptoms may achieve high reliability precisely because the symptoms are easy to identify; yet this agreement says nothing about whether the diagnosis represents a distinct psychiatric disorder.

Reliability may thus arise as an illusion of consistency — a technical agreement concerning surface-level symptoms rather than valid nosological entities. The classic analogy is straightforward: if all patients presenting with fever and headache are defined as having the same disease, the diagnosis would demonstrate high reliability but no validity, as the symptoms are nonspecific and may result from many different conditions.

As Thomas Insel remarked during the launch of DSM-5, "Diagnostic categories lack validity"14. This statement captures the culmination of a broader methodological trajectory that, since 1980, has systematically prioritized reliability over validity. In doing so, psychiatry has sought procedural reproducibility at the expense of conceptual and empirical coherence—a path that has proven epistemologically and scientifically untenable. By transforming reliability from a means of achieving validity into an end, the field has substituted statistical agreement for explanatory understanding.

This overemphasis on reliability has resulted in diagnostic constructs that are stable in form yet fragile in meaning. A system can be internally consistent without being scientifically accurate; indeed, consistency may conceal rather than correct conceptual error. As Kendler (2006)22 argues, reliability is necessary but not sufficient for validity: a diagnosis that can be repeatedly applied yet lacks causal grounding or predictive value fails to meet the basic criteria of scientific classification. The historical reliance on kappa thus reveals a deeper epistemic confusion—mistaking agreement among observers for correspondence with reality.

The Problem of Kappa Measurement in Practice

Different methods of assessing reliability have revealed discrepancies that cast doubt on the practical utility of the kappa statistic:

- Audio recording method (one interviewer, multiple raters): produces artificially high kappa values, since all raters assess the same standardized interview.
- Test–retest method (two separate interviews): yields lower kappa values, as variation in interview style and questioning exposes inconsistency in diagnostic judgments.

These discrepancies demonstrate that the closer a method approximates real-world clinical practice, the lower the reliability becomes. This undermines the usefulness of kappa as a quality metric in realistic and complex diagnostic contexts.

From Technical Reliability to Nosological Meaning

The concept of reliability in psychiatric diagnosis must therefore be reconsidered. A metric that fails to distinguish between non-specific and differentiated symptoms risks reinforcing diagnostic categories that lack nosological significance.

It is not sufficient that two clinicians agree that a patient appears restless and has difficulty concentrating. The critical question is:

Are these symptoms specific to a single diagnosis, or do they belong to a broader symptom complex shared by multiple conditions?

When reliability is measured without regard to context, etiology, or comorbidity, it loses its value as a scientific indicator.

The Limitations of Kappa and Its Problematic Relationship to Comorbidity - High Kappa Does Not Necessarily Equal High Reliability

As discussed earlier, Cohen's kappa is used as a measure of interrater reliability — the degree of agreement between clinicians assessing the same patient independently. A high kappa value is generally interpreted as evidence of high reliability. However, this assumes that the observed agreement reflects a precisely delineated and differentiated diagnostic category. This assumption represents one of the major methodological weaknesses of kappa-based reliability in psychiatry.

If a diagnosis is defined by broad, easily recognizable, and non-specific symptoms — such as restlessness, concentration difficulties, or impulsivity — clinicians are more likely to agree. Not because the diagnosis possesses high validity, but because such symptoms occur across numerous psychiatric and somatic conditions and are easy to identify.

This creates a false sense of security in the kappa coefficient: a high value may give the appearance of reliability, while in reality concealing a lack of nosological precision and limited clinical utility.

Diagnoses with High Comorbidity Create Methodological Problems

Another major problem with kappa arises when it is applied to diagnostic categories characterized by high comorbidity — that is, a strong tendency to co-occur with multiple other disorders. ADHD provides a clear example. Several meta-analyses and population studies have shown that ADHD frequently coexists with as many as 10–15 other diagnoses, including anxiety disorders, affective disorders, autism spectrum disorder, Tourette's syndrome, bipolar disorder, PTSD, sleep disturbances, and conduct disorders [23-25] (Ghanizadeh, 2012; Willcutt et al., 2012; Newson et al., 2021).

These high comorbidity rates indicate that the symptoms defining the diagnosis are neither unique nor differentiated. On the contrary, the more a diagnostic category overlaps with other disorders, the less informational value it possesses. It becomes increasingly unclear whether the diagnosis identifies a distinct psychiatric entity or merely aggregates non-specific symptoms that are already distributed across numerous other conditions.

When this overlap is combined with high kappa values — which arise precisely because clinicians recognize and agree on such broad symptoms — we encounter a situation in which reliability appears high while validity deteriorates.

I refer to this phenomenon as mimetic reliability: it imitates reliability but is, in fact, misleading.

Introduction of M-Kappa: A Correction for Misleading Reliability - From Technical Reliability to Nosological Precision

As demonstrated above, the classical kappa coefficient (Cohen's κ) quantifies interrater reliability — that is, the extent to which two or more clinicians independently agree on a diagnosis, adjusted for agreement expected by chance. However, when the diagnostic criteria are broad and based on symptoms that occur across multiple conditions, a high kappa value does not necessarily indicate diagnostic precision. On the contrary, agreement may emerge precisely because the symptoms are non-specific and easily recognizable across disorders.

In cases where a diagnosis exhibits high comorbidity — and thus shares a substantial proportion of its symptomatology with numerous other conditions — the kappa coefficient loses its interpretive value. In such situations, the diagnosis functions as a vacuum cleaner for symptoms, aggregating features that no longer distinguish it nosologically from adjacent categories.

In this context, reliability as measured by the kappa coefficient is merely a measure of agreement, not a scientifically meaningful construct. The proposed M-Kappa, a corrected reliability coefficient, seeks to restore scientific meaning to the kappa statistic by accounting for variables that artificially inflate its value. M-Kappa modifies the classical kappa (K) by incorporating the effects of comorbidity and symptom overlap, thereby providing a more valid representation of diagnostic reliability.

Theoretical Definition of M-Kappa

$$M\text{-}Kappa \cdot K(\sum_{i=1}^{n} \frac{1}{\frac{P_i - P_0}{P_0}})$$

Where:

- **K** is the observed kappa (interrater reliability).
- **n** represents the number of comorbid conditions whose prevalence exceeds the normal base rate.
- **P_0** denotes the baseline prevalence of the condition in the general population (typically 10% or lower).
- **P_i** represents the prevalence of the ith comorbid condition among individuals with the target diagnosis.

The term $(P_{i----}P_0)/P_0$ expresses the relative overrepresentation of each comorbidity. The summation calculates the average relative overrepresentation across all relevant comorbidities.

When relative overrepresentation is high, the adjusted kappa value is proportionally reduced — reflecting that the observed reliability likely arises from superficial or overlapping symptoms rather than from genuine nosological differentiation.

Interpretation of the Model

The figure above illustrates M-Kappa as a function of the comorbidity rate (R). As comorbidity increases, the corrected kappa value decreases proportionally, demonstrating that apparent diagnostic agreement may be inflated by symptom overlap rather than reflecting true reliability. In this sense, M-Kappa operationalizes a correction from technical reliability toward epistemic validity, emphasizing that agreement alone is insufficient unless it pertains to diagnostically distinct and etiologically coherent constructs.

Empirical Application of M-Kappa

The M-Kappa coefficient can be empirically applied in both clinical and research contexts to evaluate the robustness and validity of psychiatric diagnoses. In principle, it functions as a corrective adjustment to traditional reliability estimates in studies where comorbidity and symptom overlap are known to distort kappa values.

In field trials — such as those conducted during the development of DSM-5 — M-Kappa could be used retrospectively to re-evaluate diagnostic reliability scores by incorporating prevalence data on comorbid disorders. For example, if a diagnostic category demonstrates high co-occurrence with anxiety or affective disorders, M-Kappa would proportionally reduce the apparent reliability, providing a more realistic estimate of diagnostic distinctiveness. This adjustment would allow researchers to distinguish between true diagnostic consistency and mimetic reliability — the latter reflecting artificial agreement driven by overlapping symptomatology rather than by a coherent nosological entity.

Moreover, M-Kappa lends itself to implementation in machine learning and computational psychiatry, where algorithmic classification of symptoms is increasingly used to model diagnostic boundaries. By integrating comorbidity-adjusted parameters into reliability calculations, M-Kappa could serve as a validation metric for automated diagnostic systems, penalizing models that achieve high accuracy through recognition of non-specific or redundant symptom patterns. In this way, M-Kappa aligns with emerging frameworks in computational phenotyping that emphasize dimensional, data-driven validity over categorical agreement alone.

Finally, the application of M-Kappa invites a reconsideration of how reliability is conceptualized within psychiatric science. Rather than viewing reliability as a static statistical outcome, it can be understood as a dynamic epistemic construct — one that must account for the complexity of overlapping symptom networks, probabilistic associations, and shared etiological mechanisms. In doing so, M-Kappa reintroduces philosophical depth to the measurement of reliability, bridging methodological rigor with nosological meaning.

Symptom Overlap and the Illusion of Reliability

Empirical evidence further illustrates the extent to which symptom overlap can produce misleading reliability estimates. Diagnoses such as ADHD often achieve kappa values between 0.6 and 0.8 in research contexts[24]. Yet comorbidity rates in the studies typically range from 60 % to 80 % [23,25]. The diagnostic criteria themselves — for example, “interrupts others,” “leaves seat,” or “difficulty sustaining attention” — also occur across a wide range of other disorders, including anxiety, bipolar disorder, PTSD, autism spectrum disorder, and sleep disturbances [27-29.]

In a forthcoming manuscript30, it is demonstrated that ADHD symptoms overlap with all 25 psychiatric diagnoses examined—varying in extent but consistently occurring at a high level. The severity of this overlap appears to determine the degree to which ADHD-related symptoms recur across diagnostic categories, suggesting that the magnitude of overlap functions as a key indicator of diagnostic invalidation. This finding implies that even when kappa values appear high, they may conceal a diagnostic structure so symptomatically diffuse that the observed reliability becomes illusory. What appears to be agreement between clinicians may instead reflect the repeated identification of non-specific behavioral traits rather than a coherent, nosologically valid entity.

Conclusion: From Mimetic to Meaningful Reliability

The M-Kappa framework redefines reliability as a dynamic and context-sensitive measure — one that distinguishes between superficial agreement and substantive diagnostic coherence. By adjusting for comorbidity and symptom overlap, M-Kappa corrects the false inflation of traditional kappa values, revealing the extent to which high reliability can coexist with low validity.

In doing so, M-Kappa moves psychiatric methodology beyond the illusion of “mimetic reliability” — the imitation of precision produced by non-specific symptom consensus — toward a form of reliability that is epistemically and nosologically meaningful. This shift underscores a central principle: reliability without validity is not scientific precision, but systematic error.

By integrating comorbidity-adjusted parameters, M-Kappa invites a broader reconsideration of diagnostic science. It restores the conceptual balance between reproducibility and meaning, demonstrating that reliability must serve validity — not replace it. In this sense, M-Kappa represents both a methodological correction and a philosophical realignment: from imitation to understanding, from statistical agreement to scientific explanation.

Kappa was introduced as a method intended to make psychiatric research more scientific, but as shown here, it had the opposite effect. With M-kappa, we can now obtain kappa values that carry genuine scientific meaning.

References

1. American Psychiatric Association. (1980). Diagnostic and Statistical Manual of Mental Disorders (2nd ed.). APA.
2. American Psychiatric Association. (1952). Diagnostic and Statistical Manual of Mental Disorders (1nd ed.). APA.
3. American Psychiatric Association. (1968). Diagnostic and Statistical Manual of Mental Disorders (2nd ed.). APA.
4. Spitzer, R. L., & Fleiss, J. L. (1974). A re-analysis of the reliability of psychiatric diagnosis. British Journal of Psychiatry, 125(0), 341–347.
5. Spitzer R.L Values and Assumptions in the Development of DSM-III and DSM-III-R: An Insider's Perspective and a Belated Response to Sadler, Hulgus, and Agich's "On Values in Recent American Psychiatric Classification". The Journal of Nervous and Mental Disease 189(6):p 351-359, June 2001.
6. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.
7. Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. Family Medicine, 37(5), 360-363.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74.
9. Kundel HL, Polansky M. Measurement of observer agreement. Radiology. 2003 Aug;228(2):303-8.
10. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005 Mar;85(3):257-68. PMID: 15733050.
11. Spitzer RL, Endicott J, Robins E. Research diagnostic criteria. Psychopharmacol Bull. 1975 Jul;11(3):22-5.
12. Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. Archives of General Psychiatry, 26(1), 57–63.
13. Kirk, S. A., & Kutchins, H. (1992). The selling of DSM: The rhetoric of science in psychiatry. Aldine de Gruyter.
14. Greenberg, G. (2013, May 7). The NIMH withdraws support for DSM-5. The New Yorker. https://www.newyorker.com/news/news-desk/the-nimh-withdraws-support-for-dsm-5
15. Deacon, B. J. (2013). The biomedical model of mental disorder: A critical analysis of its validity, utility, and effects on psychotherapy research. Clinical Psychology Review, 33(7), 846–861. https://doi.org/10.1016/j.cpr.2012.09.007
16. Mayes, R., & Horwitz, A. V. (2005). DSM-III and the revolution in the classification of mental illness. Journal of the History of the Behavioral Sciences, 41(3), 249–267. https://doi.org/10.1002/jhbs.20103
17. Frances, A. (2013). Saving Normal: An Insider’s Revolt Against Out-of-Control Psychiatric Diagnosis, DSM-5, Big Pharma, and the Medicalization of Ordinary Life. New York: William Morrow.
18. Pickersgill M.D. Debating DSM-5: diagnosis and the sociology of critique. J Med Ethics. 2014 Aug;40(8):521-5.
19. Williams JB, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, Howes MJ, Kane J, Pope HG Jr, Rounsaville B, et al. The Structured Clinical Interview for DSM-III-R (SCID). II. Multisite test-retest reliability. Arch Gen Psychiatry. 1992 Aug;49(8):630-6.
20. Regier D.A., William E. Narrow, M.D., M.P.H.; Diana E. Clarke, Ph.D., M.Sc.; Helena C. Kraemer, Ph.D.; S. Janet Kuramoto, Ph.D., M.H.S., Emily A. Kuhl, Ph.D & David J. Kupfer, M.D ” DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest

Reliability of Selected Categorical Diagnoses". Am J Psychiatry 2013; 170:59–7

21. Freedman R., Lewis DA, Michels R, Pine DS, Schultz SK, Tamminga CA, Gabbard GO, Gau SS, Javitt DC, Oquendo MA, Shrout PE, Vieta E, Yager J. The initial field trials of DSM-5: new blooms and old thorns. Am J Psychiatry. 2013 Jan;170(1):1-5
22. Kendler, K. S. (2006). The nature of psychiatric disorders. World Psychiatry, 5(3), 167–171.
23. Ghanizadeh, A. (2012). Comorbidity of ADHD and anxiety disorders: A systematic review. Research in Developmental Disabilities, 33(6), 2301–2308.
24. Willcutt, E. G., Nigg, J. T., Pennington, B. F., Solanto, M. V., Rohde, L. A., Tannock, R., ... & Lahey, B. B. (2012). Validity of DSM-IV attention-deficit/hyperactivity disorder symptom dimensions and subtypes. Journal of Abnormal Psychology, 121(4), 991–1010.
25. Newson JJ, Pastukh V, Thiagarajan TC. Poor Separation of Clinical Symptom Profiles by DSM-5 Disorder Criteria. Front Psychiatry. 2021 Nov 29;12:775762.
26. Gillberg, C., Gillberg, I. C., Rasmussen, P., Kadesjö, B., Söderström, H., Råstam, M., Johnson, M., Rothenberger, A., & Niklasson, L. (2003). Co-existing disorders in ADHD—Implications for diagnosis and intervention. European Child & Adolescent Psychiatry, 13(1), 80–92.
27. Barrett, P. M., Dadds, M. R., & Rapee, R. M. (1996). Family treatment of childhood anxiety: A controlled trial. Journal of Consulting and Clinical Psychology, 64(2), 333–342.
28. Mostofsky, S. H., & Ewen, J. B. (2011). Altered connectivity and action model formation in autism spectrum disorders. Physiology & Behavior, 104(2), 354–359.
29. van der Kolk, B. A. (2014). The body keeps the score: Brain, mind, and body in the healing of trauma. Viking.
30. von Bergen, C. (in preparation). Symptom overlap in ADHD diagnosis: A cross-nosological analysis.