Validation strategies for subtypes in psychiatry: A systematic review of research on autism spectrum disorder

Agelink van Rentergem, J.A.; Deserno, M.K.; Geurts, H.M.

DOI
10.1016/j.cpr.2021.102033

Publication date
2021

Document Version
Final published version

Published in
Clinical Psychology Review

License
CC BY

Link to publication

Citation for published version (APA):
Agelink van Rentergem, J. A., Deserno, M. K., & Geurts, H. M. (2021). Validation strategies for subtypes in psychiatry: A systematic review of research on autism spectrum disorder. *Clinical Psychology Review*, *87*, [102033]. https://doi.org/10.1016/j.cpr.2021.102033

Contents lists available at ScienceDirect

Clinical Psychology Review

journal homepage: www.elsevier.com/locate/clinpsychrev

Review

Validation strategies for subtypes in psychiatry: A systematic review of research on autism spectrum disorder

Joost A. Agelink van Rentergem [a,b,*], Marie K. Deserno [a,b], Hilde M. Geurts [a,b,c]

[a] Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands
[b] Dutch Autism & ADHD Research Center, the Netherlands
[c] Dr. Leo Kannerhuis, the Netherlands

ABSTRACT

Heterogeneity within autism spectrum disorder (ASD) is recognized as a challenge to both biological and psychological research, as well as clinical practice. To reduce unexplained heterogeneity, subtyping techniques are often used to establish more homogeneous subtypes based on metrics of similarity and dissimilarity between people. We review the ASD literature to create a systematic overview of the subtyping procedures and subtype validation techniques that are used in this field. We conducted a systematic review of 156 articles (2001-June 2020) that subtyped participants (range N of studies = 17–20,658), of which some or all had an ASD diagnosis. We found a large diversity in (parametric and non-parametric) methods and (biological, psychological, demographic) variables used to establish subtypes. The majority of studies validated their subtype results using variables that were measured concurrently, but were not included in the subtyping procedure. Other investigations into subtypes' validity were rarer. In order to advance clinical research and the theoretical and clinical usefulness of identified subtypes, we propose a structured approach and present the SUbtyping VAlidation Checklist (SUVAC), a checklist for validating subtyping results.

1. Introduction

The characteristics of people diagnosed with Autism Spectrum Disorder (ASD) vary greatly, even though ASD is characterized by challenges in social interactions and communication, restrictive repetitive behaviors, and sensory sensitivities (American Psychiatric Association, 2013). Originally, a narrow category of children received an ASD diagnosis, when they were severely impaired in their social and communication skills, and could hardly bear changes in their environment (Kanner, 1943). Over the years, diagnostic criteria have changed, and now include a much wider spectrum (Wing & Potter, 2002). The prevalence has increased, from 4 cases in 10,000, to around one case in 100 (Elsabbagh et al., 2012; Fombonne, 2018), which is most likely due to widening of the criteria (Mottron & Bzdok, 2020) and an increase in recognition, rather than an increase in actual incidence, as ASD prevalence is stable across different ages (Brugha et al., 2011). For example, while ASD used to be diagnosed primarily in boys of European descent, there are many developments in increased diagnosis of ASD in girls (Lai, Lombardo, Auyeung, Chakrabarti, & Baron-Cohen, 2015), adults and elderly (Piven & Rabins, 2011), and non-Caucasian populations (Becerra et al., 2014). When more people qualify to receive an ASD diagnosis, the group of people with an ASD diagnosis will become more heterogeneous (Mottron & Bzdok, 2020). Increasing heterogeneity of the ASD population will come with even more difficulties to formulate straightforward clinical advice within support programmes. Heterogeneity between people with an ASD diagnosis is already causing difficulties in finding causes and interventions for this population (Happé, Ronald, & Plomin, 2006). In the current article we review the literature on subtyping people with a diagnosis of ASD, and specifically focus on what validation strategies researchers use to make sure that their subtypes are useful, reliable, and valid.

In the scientific (ASD) literature, the term heterogeneity is used in various ways. Some use the word heterogeneity to describe random variability between individuals (e.g., Georgiades, Szatmari, & Boyle, 2013). People vary in psychological traits (e.g., in personality or ability) and in biological characteristics (e.g., gene expression, brain morphology). For example, on two questionnaires, 50 people with an ASD diagnosis may obtain 50 different combinations of scores. Such random variability also complicates the search for causes, as a cause would be more readily identified if all people with an ASD diagnosis were identical. However, random variability is not the kind of heterogeneity we refer to here.

We define heterogeneity as the existence of subtypes that are qualitatively different. This can be in the psychological or biological domain, or a combination of both. For example, 50 people with an ASD diagnosis who fill in two questionnaires may form two subtypes, with 30 people

[*] Corresponding author at: Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 WS Amsterdam, the Netherlands.
E-mail address: j.a.agelink@uva.nl (J.A. Agelink van Rentergem).

https://doi.org/10.1016/j.cpr.2021.102033
Received 19 June 2020; Received in revised form 14 February 2021; Accepted 14 April 2021
Available online 19 April 2021

obtaining a high score on the first questionnaire and a low score on the second, and 20 people obtaining a low score on the first questionnaire and a high score on the second. For the first subtype, genetic causes may be responsible for their difficulties, while environmental causes may be responsible for the second subtype's difficulties. These causes become much harder to identify without knowing that subtypes exist, and without knowing to which subtype people belong. This illustrates the importance of identifying valid subtypes, as a possible prerequisite for identifying causes.

In the ASD research realm, various attempts have been undertaken to tackle heterogeneity by establishing more homogeneous subtypes, to meet specific needs of specific subtypes. There are many reasons that subtyping analyses are desirable (Grzadzinski, Huerta, & Lord, 2013, Georgiades et al., 2013). First, if we can assign people with a high degree of certainty to subtypes, we can study what the prognosis is for people in different subtypes, and provide better information to people on what to expect later in life (Bohane, Maguire, & Richardson, 2017). Second, if a subtype can be identified that is homogeneous in the constellation of behaviors that people show, this could aid the search for biomarkers for these behaviors. If such biomarkers exist, these can be used in early diagnosis, and therefore early interventions, potentially leading to better outcomes. Differences in biomarkers between subtypes could also be the cause of subtype membership. Third, if we can assign people to subtypes, we can find out what kind of intervention works best for which subtype, and which intervention may even be disadvantageous for a particular subtype.

Prognosis, predictors of subtype membership and heterogeneity of intervention effects all relate to outcomes that are external to the subtypes themselves. This focus on the predictive value of a subtyping result is in line with recent recommendations on tying subtyping methods to predictive methods. Such methods can ensure that subtyping results will also have practical implications (Feczko et al., 2019), given that there are theoretical or clinically motivated reasons that there should be a relationship between the outcome to be predicted, and the constellation of variables used to form subtypes.

There are also ontological reasons to study subtypes, which are of intrinsic value regardless of external outcomes. A subtyping analysis can examine whether individual differences reflect subtypes, or individual differences reflect a dimension (Bernstein et al., 2010). A dimension on which people differ randomly may cause similar problems as undetected subtypes, and may cause researchers to presuppose the existence of subtypes (Widiger, 1992). A subtyping analysis would be required to discover the absence of subtypes. Second, subtyping analyses can examine established delineations between disorders. For example, delineations between ASD and conditions like schizophrenia and ADHD (Eack et al., 2013) can be studied, to examine whether the current delineations are optimal in assigning people to the best possible intervention, or whether an alternative delineation may better represent individual differences in the associated psychopathology. If a single diagnosis is found to be a combination of multiple subtypes, this could then lead to an evidence-based split of categories in diagnostic manuals (Brewin et al., 2017).

1.1. Subtyping in ASD

In this article, we review the literature on empirical subtyping of people with ASD. To our knowledge, there have been five past reviews of subtyping in ASD (Beglinger & Smith, 2001; DeBoth & Reynolds, 2017; Marquand, Wolfers, Mennes, Buitelaar, & Beckmann, 2016; Syriopoulou-Delli & Papaefstathiou, 2020; Wolfers et al., 2019). Beglinger and Smith (2001) provide an overview of the literature up to 2001, and include 17 different studies on subtypes of ASD. Although there are some discrepant results, their review of the evidence suggests that there are around four different subtypes that can be discerned in every study, with differences in results depending on what variables are included in the subtyping method. Generally, most results indicate a severity gradient, with subtypes that are ordered in the sense that one subtype is least affected, and one is most affected across different variables. DeBoth and Reynolds (2017) provided an overview of the literature on subtyping of people with an ASD diagnosis on the basis of sensory-based measures, and included eight articles. Their review of the evidence indicates that generally, there are three to five subtypes, depending on whether measures of both hyporeactivity and hyperreactivity were included. Marquand et al. (2016) provided a review of the literature on subtyping in psychiatry in a broader sense. For ASD, they discuss six recent articles, and highlight the diversity in used variables and resulting subtypes. Recently, Wolfers et al. (2019) provided a review of the literature on subtyping in ASD. In comparison to the present article, they included fewer articles that are relevant to this discussion (19, vs. 156 articles included in the present article). This difference in article inclusion is most probably due to the authors' decision to include a shorter time frame, use less comprehensive search terms, and restrict the search to a single database. Furthermore, their review recorded two validation strategies, while the present review distinguishes seven. Similarly, the Syriopoulou-Delli and Papaefstathiou (2020) review included an even smaller number of articles (10 articles).

One aspect of subtyping analyses that is understudied in each of these reviews is the way in which results are substantiated, which we call validation strategies. If an analysis finds four different subtypes, there is little information to either corroborate or contest the existence of those four subtypes, and it remains an open question whether the four subtypes are a chance finding. There is some implicit information in the methodological rigor of the study design, and the number of data points that were collected. However, a subtyping result in itself does not provide information on whether the results are generalizable to the broader population, replicable in other research, or useful to other researchers and clinicians in their thinking.

In the present article, the current state of validation in the literature on empirical subtyping studies in ASD is therefore reviewed. Since the review of Beglinger and Smith (2001), many articles have been published beyond the 17 that they included, using a wide variety of samples, variables, methods, and ways of validating the results. Also, we focus on empirical subtyping methods, excluding articles that use preset cutoffs to form subtypes. In contrast to the review of DeBoth and Reynolds (2017) that focused on sensory variables, our review does not focus on a single domain, but considers all types of variables that have been used to find subtypes in ASD. We systematically review the literature between 2001 and June 2020, including every study that uses an empirical method to find subtypes within a sample of autistic people. In contrast to Wolfers et al. (2019), we look at a more representative sample of articles that are relevant to this discussion, and focus on a wider range of validation measures. The main question we aim to answer is: Does the growing body of subtyping literature provide sufficient corroborating evidence to suggest valid and reliable subtypes? And if we can answer this question affirmatively, what are the subtypes that are most consistently supported by the literature?

1.2. Validation strategies for subtypes

In the rest of the review, we aim to identify applications of seven validation strategies in the literature, defined as follows. The first two validation strategies —"cross-method replication" and "subtype separation"— can be applied with a single data set, with a single set of measures. These strategies are depicted in Fig. 1. In "cross-method replication", subtypes are formed with two or more statistical methods, comparing the results. For example, hierarchical clustering and k-means clustering techniques can be applied to a single dataset, to see whether each technique results in the same number and score profile of subtypes. For example, in one study, four different methods were applied to a single dataset to establish whether the number of subtypes was stable across methods (Hu & Steinberg, 2009). The reasoning is that subtypes are clearly distinguishable when they can be detected with disparate

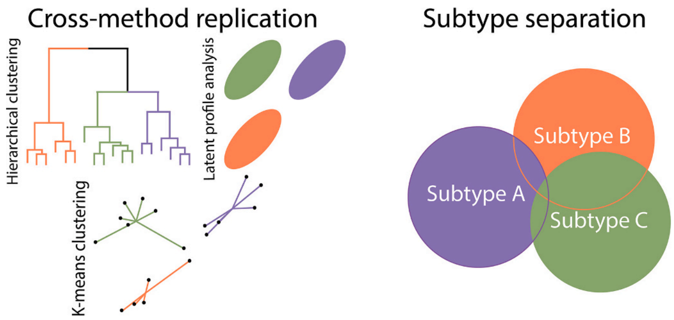

Fig. 1. Illustration of the cross-method replication and subtype separation validation strategies.

statistical methods (Taylor, Asmundson, & Carleton, 2006).

One metric for the validity of subtypes is the certainty with which participants are assigned to different subtypes. We refer to this as "subtype separation", as it measures whether the subtypes are clearly separated and distinct, or whether there is overlap between subtypes. For these purposes, statisticians have developed indices like the "mean posterior probability of class membership", which quantifies the degree of certainty with which people are assigned to specific subtypes. The reasoning is that subtypes are more valid if people are consistently a member of one, and only one, subtype, rather than being a possible member of multiple subtypes.

The third and fourth validation strategies —"independent replication" and "temporal stability"— require extra data collection using the same measures, either testing new participants or testing participants a second time. These strategies are depicted in Fig. 2. In "independent replication", subtypes are constructed based on two different samples, using the same measures. The two samples are independent, and the initial subtyping result can be replicated. For example, a sample can be split into two, validating the results of the analysis of the first half on the second half. In one study with an intellectually disabled sample, the eight-subtype solution that was found for the first sample was validated in a second sample (Brown, Aman, & Lecavalier, 2004). The reasoning is that if subtypes exist in the population, analysis of any representative sample from this population should recover the same number and type of subtypes.

In "temporal stability", subtypes are formed at one measurement occasion, and established again at a later measurement occasion, using the same measures. We want to know whether subtype membership is stable over time, or whether participants switch between subtypes. To establish stability, participants can be retested after a number of years, and subtypes can be constructed once more with this data, comparing the results of the two subtyping analyses. For example, in one study, children with an ASD diagnosis that were subtyped at the time of diagnosis were retested and re-analyzed at age 6 with the same subtyping technique, to find that the children in the three subtypes at baseline were now divided over two subtypes, which did not correspond one-to-one with one of the subtypes at baseline (Georgiades et al., 2014). The reasoning is that if subtypes are valid to the point where we can find causal biomarkers for them, the number of subtypes should remain the same, and subtype membership should not vary too much over time. In a review of subtypes in the eating disorder literature, a consistent three subtypes were found across studies, but the complete absence of investigations into temporal stability was identified as an important limitation throughout (Wildes & Marcus, 2013).

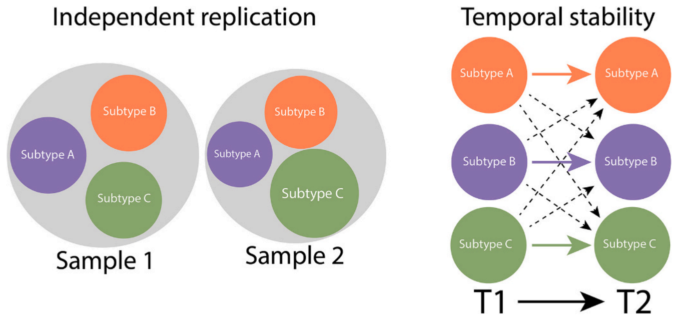

Fig. 2. Illustration of the independent replication and temporal stability subtype validation strategies.

The final three validation strategies —"external validation", "parallel validation", and "predictive validation"— require that data on more variables are collected, outside of the variables that are used in the subtyping procedure. These strategies are depicted in Fig. 3. In "external validation", subtypes are compared on variables that were not used in the construction of the subtypes, and that are theorized to be related to interindividual differences. For example, subtypes can be compared on demographic variables or other variables that should theoretically be different between subtypes, but were not used in the construction of the subtypes. This was done with subtypes constructed using age, cognitive abilities, and adaptive functioning, after which the subtypes were compared on the scores they obtained on a checklist of ASD behaviors (Bitsika, Sharpley, & Orapeleng, 2008). The reasoning is that differences between subtypes should not be limited to variables used to construct the subtypes.

In "parallel validation", subtypes are constructed with the same sample at the same measurement occasion, with different variables that are theoretically equivalent to the variables that are used in the subtyping. For example, latent trajectory subtypes were found to be the same in a longitudinal study of children with an ASD diagnosis, regardless of which measure of daily living skills was used (Bal, Kim, Cheong, & Lord, 2015). The reasoning is that this would indicate that not the chosen measurement instruments themselves are important in determining subtypes, but that the constructs that underlie the measurement instruments are important.

In "predictive validation", subtype membership is used to predict variables on a later measurement occasion. This is similar to both "external validation" and "temporal stability", as information is used on other variables than are used to form subtypes, and data is used from a later measurement occasion. If subtypes are found to differ on variables at a later measurement occasion, this is evidence that the subtypes are not only distinct, but also have prognostic value for the individual. The reasoning is that if subtypes are found to provide reliable predictions for future outcomes, this means that they are not only valid in the sense of describing real differences between subtypes, but are also clinically relevant. As noted in a review of OCD subtypes, using subtypes to predict treatment response is done in relatively few studies, even though also in OCD, results suggest that treatments need to be adjusted to the specific subtype (McKay et al., 2004).

The application of these validation strategies is far from identical across studies. Different studies use different indices to establish "subtype separation". Also, studies do not necessarily use the term "external validation" to describe comparisons between subtypes on additional outcome measures. However, all validation strategies that are used in the literature to corroborate the existence of subtypes can be classified as belonging to one of these seven.

2. Methods

2.1. Literature

2.1.1. Search strategy

The literature search strategy combined keywords related to ASD diagnoses (variations of autism, Asperger's and Pervasive Developmental Disorder) with keywords related to the different types of subtyping methods (exact search syntax in Appendix): parametric methods (variations of latent class analysis, mixture models, etc.), non-

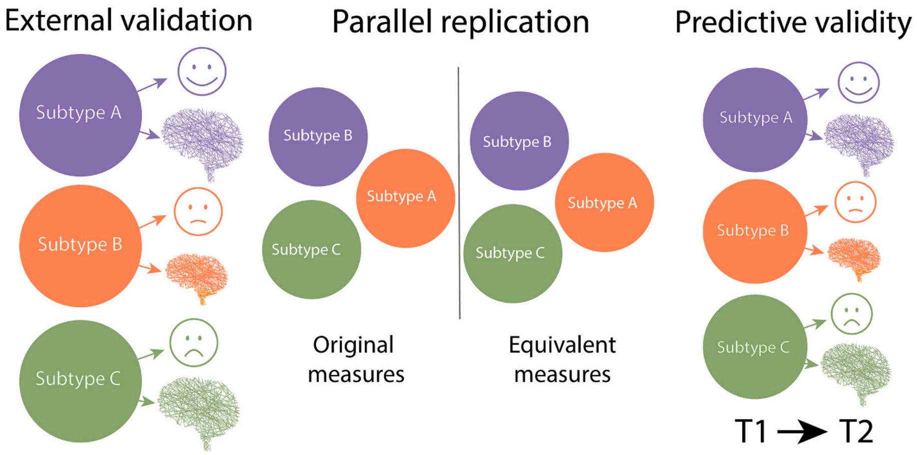

Fig. 3. Illustration of the external validation, parallel replication, and predictive validity validation strategies. For external validation, emotional and brain outcomes would not be included in the formation of the subtypes. For predictive validation, outcomes would not be included in the formation of subtypes, and would be measured at a later measurement occasion.

parametric methods (variations of k-means, hierarchical clustering, etc.) and community detection methods (variations of community detection, cliques). Both PsycINFO and MEDLINE (on which PubMed is based) databases were searched, because these cover different portions of the literature (Wu, Aylward, Roberts, & Evans, 2012). In all articles, references to other subtyping analyses were inspected to make sure they were included if they were not found by the initial search.

2.1.2. *Inclusion and exclusion criteria*

There were seven inclusion criteria, relating to publication date, samples, measures, and analyses. The first inclusion criterion was publication after 2000. Two searches were conducted. The first was conducted in February 2018, and included papers published between January 2001 and February 2018. The second was an update in June 2020, and included papers published between February 2018 and June 2020. The second inclusion criterion was that living humans were studied as test subjects. The third inclusion criterion was that at least part of the sample had an ASD diagnosis.

The fourth inclusion criterion was that measurements were taken that related to the person with the ASD diagnosis. This for example excluded studies that measured the behavior of mothers of children with an ASD diagnosis. However, proxy ratings were included, i.e., ratings that mothers provided of the behavior of their children with an ASD diagnosis. The fifth inclusion criterion was that the subtyping method was used to assign people to subtypes. The sixth inclusion criterion was that an empirical statistical method of subtyping was used to assign people. This excluded studies that used predefined subtype descriptions to assign people, which might have been established on theoretical grounds or on earlier empirical work. Articles that featured only taxometric analyses (Bernstein et al., 2007; Meehl, 1995) —aimed at identifying whether there are two subtypes or no subtypes— were also not included. The seventh inclusion criterion was that an unsupervised method was used, i.e., a method that finds subtypes rather than a method to predict a particular outcome. This excluded, for example, support vector machines, and other classifiers. Studies were included that find novel approaches to adapt existing supervised learning methods to the unsupervised case.

2.2. *Data recording*

For the first main search, all data were recorded by one of two authors (JAR, MKD), with each checking the other's coding. Data for the update were recorded by the first author (JAR). Furthermore, a number of checks were performed, correcting any possible errors (e.g., 80% being coded as 0.8). Aside from article characteristics, like authors and publication date, we recorded data from each article on four levels: Sample characteristics, variable characteristics, analysis characteristics, and validation characteristics. The choice which data to record was based on earlier reviews of subtyping analyses (van Rooden et al., 2010, Beglinger & Smith, 2001, Marquand et al., 2016, DeBoth & Reynolds, 2017).

2.2.1. *Sample characteristics*

First, we recorded aspects of the sample that was used in the subtyping analysis. If the initial sample was larger than the sample that was analyzed in the subtyping analysis, we recorded the characteristics for the analyzed sample. We recorded the sample size, the percentage of males, the mean age and age range, the mean IQ and IQ range,[1] the percentage of participants with an ASD diagnosis and the diagnostic manual the ASD diagnosis was based on. We recorded sample sizes because they can be influential in how many subtypes are found, and how precise the delineation of different subtypes is. How large samples need to be to detect subtypes is understudied (Dziak, Lanza, & Tan, 2014). We recorded the mean age and age range of the participants because of possible differences in subtyping between infants, children, adolescents, adults, and older adults. Some studies included a broader age range, spanning multiple developmental categories. In such studies, it is of interest to see whether the subtypes that were found do not simply reflect heterogeneity in developmental stage.

Lastly, we logged the percentage of the sample with an ASD diagnosis. Some studies might have included both a typically developing group, and an ASD group. Other studies might have included an ASD group, and a group with a different diagnosis, such as schizophrenia or ADHD. Other studies only included an ASD group. If there were multiple groups in the study, but only the ASD group was used in the subtyping analysis, we only recorded the ASD group.

2.2.2. *Variable characteristics*

We recorded the number of variables that were included in the subtyping analysis, which might be the number of questionnaires, the number of subscales, or the number of items. We also documented the type of variables. There are many different kinds of variables one can use to make subtypes that can be broadly categorized as demographic, psychological, and biological. Demographic variables may for example be age, sex, and level of education. Psychological variables may for

[1] Information on IQ and diagnostic manual was often missing. These characteristics are recorded in the Table of study characteristics, but are not discussed further in the results.

example be questionnaires, cognitive tests (McCrimmon, Schwean, Saklofske, Montgomery, & Brady, 2012), or symptom checklists (Klopper, Testa, Pantelis, & Skafidas, 2017). Biological variables may for example be gene expression measurements (Kong et al., 2013), facial features (Obafemi-Ajayi et al., 2015), or EEG measures (Hasenstab, Sugar, Telesca, Jeste, & Şentürk, 2016).

2.2.3. Analysis characteristics

We recorded the type of statistical subtyping procedure, the number of subtypes that were obtained, and the relative sizes of the different subtypes in percentages of the total sample, sorted from largest to smallest.

2.2.4. Validation characteristics

We recorded whether the seven validation procedures described in the introduction were followed. To determine whether "cross-method replication" was assessed, we logged whether multiple statistical subtyping methods were used to arrive at subtypes. To determine whether "subtype separation" was assessed, we recorded whether standardized metrics were computed that quantified how distinct subtypes were, or whether the posterior probabilities of subtype membership for the participants were computed. Standardized metrics are for example the Silhouette, Dunn, and Calinski-Harabasz indices. These metrices indicate whether the variation between subtypes is large in comparison to the variation within subtypes, which reflects how separable or differentiable subtypes are. Posterior probabilities of subtype membership also reflect how separable subtypes are: If every participant can be assigned to a particular subtype with a high probability, then subtypes are more distinct than when participants can possibly belong to two or more subtypes (Nagin, 2005). Posterior membership probabilities are not available for the traditional non-parametric subtyping methods.

To determine whether an "independent replication" was undertaken, we recorded whether the subtyping result was evaluated on a sample different from the one used to establish the subtype result. This could also have been done in a cross-validation setup, where the fitting sample (the "training set" in machine learning terms) and the evaluation sample (the "test set" in machine learning terms) switch roles. To determine whether "temporal validity" was assessed, we documented whether the subtyping analysis was performed at multiple measurement occasions, for all articles that had data on multiple measurement occasions. Latent transition analysis falls within this category, as subtypes are formed at two occasions and transitions between subtypes are modeled. We did not record latent growth curve analysis as assessing temporal stability, as it uses data from multiple measurement occasions once to form subtypes, which does not convey information on stability of subtype membership over time.

To determine whether "external validity" was assessed, we logged whether subtypes were subsequently compared on variables that were not used to define the subtypes. To determine whether "parallel validation" was assessed, we recorded whether a subtype analysis was run twice in the same article, with different variables. If the variables were not clearly in different domains as considered by the authors, we recorded this as an assessment of parallel validation. To determine whether "predictive validation" was assessed, we recorded whether subtypes were compared on variables that were not used to define the subtypes, like in external validation, but that were also measured at a later measurement occasion.

For each of the validation methods, we did not record to which degree the results were valid: This is a subjective decision, and depends on the context. Therefore, our goal was to record whether these steps towards validation of the results were taken, without judging whether they were successful.

3. Results

In Fig. 4, the PRISMA diagram is provided (final $n = 156$). The records that were initially not found in our search were identified in reference lists of other articles.

In total, the samples were not always completely independent between articles, as some articles extended a sample that was collected before, some articles performed a different analysis on the same sample, and some articles added an aspect or variable to an earlier subtyping

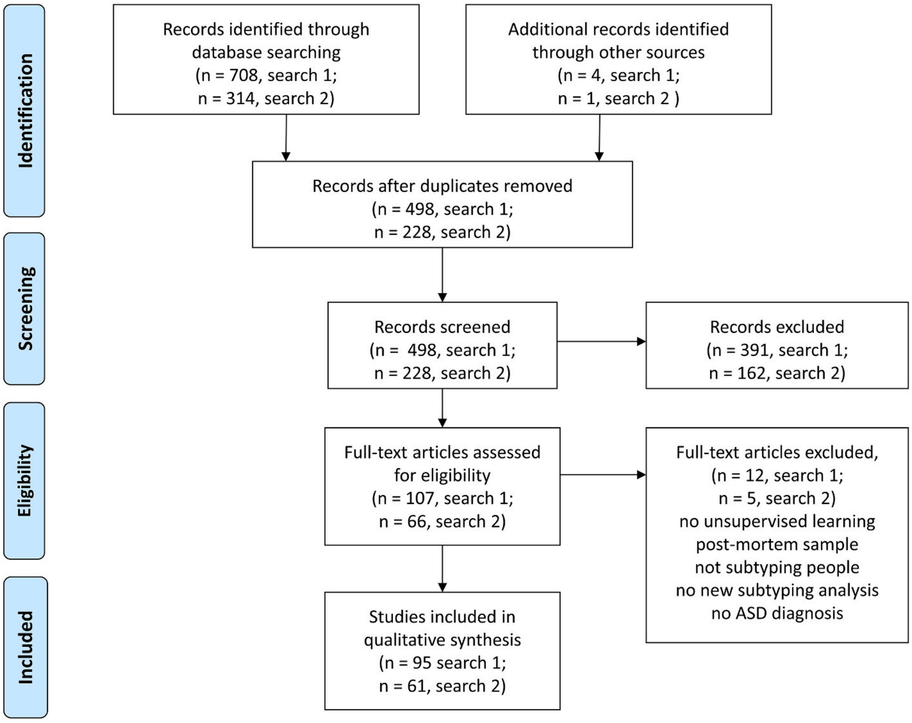
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 708, search 1; n = 314, search 2)
Additional records identified through other sources (n = 4, search 1; n = 1, search 2)
Records after duplicates removed (n = 498, search 1; n = 228, search 2)
Records screened (n = 498, search 1; n = 228, search 2)
Records excluded (n = 391, search 1; n = 162, search 2)
Full-text articles assessed for eligibility (n = 107, search 1; n = 66, search 2)
Full-text articles excluded, (n = 12, search 1; n = 5, search 2) no unsupervised learning post-mortem sample not subtyping people no new subtyping analysis no ASD diagnosis
Studies included in qualitative synthesis (n = 95 search 1; n = 61, search 2)

Fig. 4. PRISMA diagram. Search 1 was conducted in February 2018, and included papers published between January 2001 and February 2018. Search 2 was an update in June 2020, and included papers published between February 2018 and June 2020.

analysis to answer new research questions. Therefore, the 156 articles that were reported on here do not correspond to 156 unique datasets. We excluded five articles that described a subtyping analysis that had already been performed with the same sample in a different article.

The majority of the articles that we included were recent, as half of the articles were published after 2016. The number of articles that meet our criteria has been steadily increasing (Fig. 5).

3.1. Results sample characteristics

There are large differences between studies in sample size, demographics and inclusion criteria. A brief summary of each of these aspects is given below (see Table in the Appendix for study details).

3.1.1. Sample size varies from tens to tens of thousands

The median sample size was 190. Sample size ranged from $N = 17$ adults for a pilot study of language skills in adults with ASD (Lewis, Woodyatt, & Murdoch, 2008) to $N = 20,658$ for an analysis of electronic health records (Lingren et al., 2016). 32% of the samples was smaller than $N = 100$, 30% of the samples was between N = 100 and $N = 300$, and 38% was larger than $N = 300$. 15% of the studies was smaller than $N = 50$; 16% of the studies was larger than $N = 1000$. The sample size is somewhat increasing over the time frame included in this study, although studies with fewer than 100 participants remain common, see Fig. 6.

3.1.2. The majority of the participants were male

The median percentage male was 80%, with half of the studies having a percentage of males between 73% and 87%. This indicates that the inclusion rate of women into studies of subtyping was in line with current estimates of the proportion of women with an ASD diagnosis (Lai et al., 2015; note that also population studies with a minority with an ASD diagnosis were included, see below). One study studied only women (Pohl, Cassidy, Auyeung, & Baron-Cohen, 2014).

3.1.3. Most studies were conducted with child samples

The median mean age was 9, with a minimum mean age of 1.6 (Henry, Farmer, Manwaring, Swineford, & Thurm, 2018), and a maximum mean age of 45.3 (Agelink van Rentergem, Lever, & Geurts, 2019). A total of 89% of studies focused on children, defined as a mean age lower than 18. It should be noted that we recorded the age at a single measurement occasion in longitudinal studies, which should not affect the percentage, as the longitudinal studies were conducted in developing children (e.g., Bal et al., 2015; Farmer, Swineford, Swedo, & Thurm, 2018; Pickles, Anderson, & Lord, 2014; Venker, Ray-Subramanian, Bolt, & Weismer, 2014). The age range was, inevitably, wider for studies in adults. The largest age range was 2–83 years (Morris et al., 2016). The youngest tested participants were 6 months (Landa, Gross, Stuart, & Bauman, 2012; Malvy et al., 2004), the oldest 90 years (Painter, Ingham, Trevithick, Hastings, & Roy, 2018).

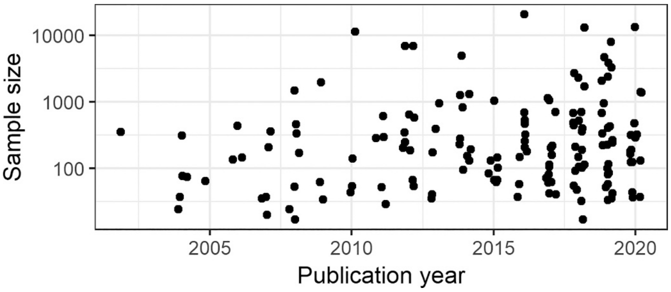

Fig. 6. Sample size by publication year. Note that the y-axis is on a log10 scale. Some random noise is added to publication year to prevent overlapping points.

3.1.4. The majority of subtyping analyses were in an all-ASD sample

For 63% of the analyses, all of the participants were diagnosed with ASD. Here, it should be noted that in studies with a mixed sample, inclusion into the subtyping analysis was leading. So, the articles coded as having a sample of whom 100% were diagnosed with ASD might have included comparison participants that were not included in the subtyping analysis. In 18% of the studies, less than half of the participants were diagnosed with ASD. Outliers were the studies where only 3–5% of the participants were diagnosed with ASD (McChesney & Toseeb, 2018, Nishimura, Takei, & Tsuchiya, 2019, Painter et al., 2018, Berlin, Lobato, Pinkos, Cerezo, & LeLeiko, 2011, Dyck, Piek, & Patrick, 2011). One of these (Berlin et al., 2011) reported ASD diagnosis status only for the subtype named “ASD”. For 4% of the studies, percentage ASD diagnosis was missing. Most studies studied either only an ASD-diagnosed group, or an ASD-diagnosed group together with a typical comparison group. However, there were studies that had looked at diagnostic boundaries between ADHD and ASD (Dajani, Llabre, Nebel, Mostofsky, & Uddin, 2016; Rommelse, van der Meer, Hartman, & Buitelaar, 2016; van der Meer et al., 2012). Another study looked at children with Down syndrome, of which some had a comorbid ASD diagnosis (Ji, Capone, & Kaufmann, 2011). A number of articles studied a diagnostically diverse sample (Castro & Pinto, 2015; Lecavalier, 2006; Little, Dean, Tomchek, & Dunn, 2017). Also, in the literature from the period where there were still divisions in the DSM between disorders that fall under ASD, subtyping analysis was used to assess whether these divisions were valid (Verté et al., 2006).

3.2. Results variable characteristics

3.2.1. Fewer than 20 variables are commonly used to construct subtypes

The median number of variables that were included in the subtyping

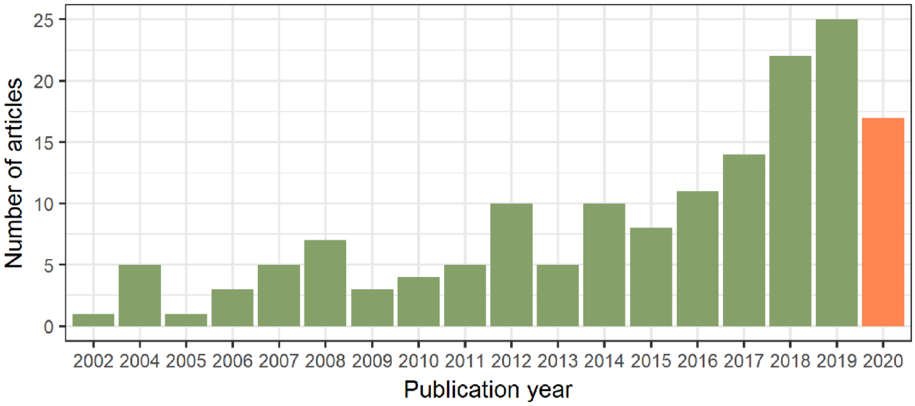

Fig. 5. Number of articles included per publication year. In orange, the number of articles published until June 2020 is plotted, so the data for 2020 is incomplete.

analysis was eight. 80% of studies included fewer than 20 variables. Larger numbers occurred sporadically, with 1350 as an absolute outlier (counts of ICD-codes over time; Doshi-Velez, Ge, & Kohane, 2014). Exceptions were a number of articles that performed latent class growth curve analyses, which typically focused on the progression on a single variable over time. These studies made up the majority of the 7% of the studies that only examined a single variable.

3.2.2. ASD characteristics are most frequently used to construct subtypes

The most frequently used variable for subtyping analyses were the Autism Diagnostic Interview – Review (ADI-R) and Autism Diagnostic Observation Schedule (ADOS). The ADI-R was used in 20 studies, although studies differed in whether subscale scores or individual items were used. The ADOS was used in 20 studies, but studies differed in whether multiple variables were used, or only a Calibrated Severity Score was entered into the analysis. Because the studies were predominantly children studies, the Vineland Adaptive Behavior Scales (VABS; 18 studies), Mullen Scales of Early Learning (10 studies) and Child Behavior Checklist (CBCL; 9 studies) were other popular choices. In total, 14 studies used variables related to sensory input (most already well-described in the specialized review mentioned in the introduction, DeBoth & Reynolds, 2017). There was generally a large diversity of variables that were used, both biological and psychological, with almost all studies having a unique set of variables included in the subtyping analysis.

3.3. Results analysis characteristics

3.3.1. Latent class analysis and hierarchical clustering are most popular

Hierarchical clustering was the most popular non-parametric method, used in 34% of the papers. k-means clustering was used in 17% of the papers. Latent Class Analysis was the most popular among the parametric methods, used in 36% of the papers. Note that under Latent Class Analysis, we subsume Latent Profile Analysis, which is applied in case of continuous measures.

A number of methods were hybrids of earlier developed methods or were otherwise too novel to fit in with the standard methods. For example, an ensemble of three methods was used in one article (Shen, Lee, Holden, & Shatkay, 2007). These five studies were coded as “Other”. Three studies made use of Two-Step Clustering, a method that is specific to the SPSS software package. Factor mixture models were used in eight studies to answer the question whether individual differences could best be described by a number of subtypes, a dimension, or a number of subtypes within which individual differences could best be described by a dimension. Latent transition analysis, a method that is especially suited for stability analyses, was performed in three studies. Latent growth curve models were used in ten studies, particularly in young children. Multivariate latent growth curve models form a theoretically strong model, and were used in three studies.

3.3.2. Two to four subtypes are recovered in the vast majority of articles

The median number of subtypes was three, and 82% of all results indicated between two and four subtypes. 11% found five, 3% found six. The largest number of subtypes was 16 (Stevens et al., 2019); these 16 subtypes were again analyzed to recover five higher-order subtypes. The lowest number of subtypes was one, which occurred in two articles (Kamp-Becker et al., 2010, Beauchamp, Rezzonico, & MacLeod, 2020), but was only one result among many analysis results in both cases. See Fig. 7 for the full distribution.

3.3.3. Substantive conclusions across articles are complicated by study heterogeneity

Because of the many differences between measures, participants' ages, sample compositions and diagnostic processes we discovered in the sample of studies, it seems premature to combine findings from studies. Subsets of studies that are more similar in terms of measures and samples become too small to draw strong conclusions. However, with the studies that have used the most popular measures (ADI-R, VABS, ADOS), we can get some impression of the stability of subtypes across more homogeneous sets of studies.

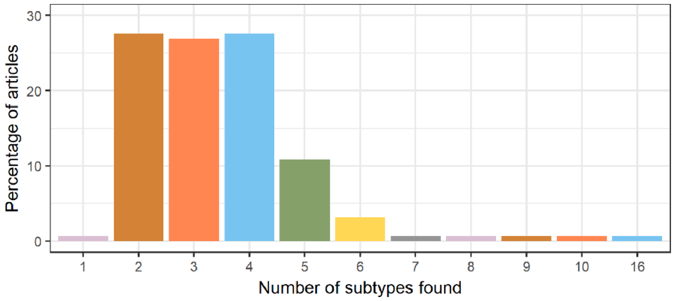

Fig. 7. Percentage of articles by number of subtypes recovered in the subtyping analysis.

Seven different studies used variables from the ADI-R in child samples where 100% had a diagnosis of ASD (Bureau, Labbe, Croteau, & Mérette, 2008; Cholemkery, Medda, Lempp, & Freitag, 2016; Georgiades et al., 2014; Hu & Steinberg, 2009; Pichitpunpong et al., 2019; Shen et al., 2007; Verté et al., 2006). Across these studies, between two and five subtypes are retrieved, mirroring the results of the entire sample of studies. It is difficult to understand where the differences in number of subtypes come from —the number of subtypes seems unrelated to publication year, statistical method, and number of variables included—and the number of studies becomes too low to further stratify these studies. Across three of the studies (Cholemkery et al., 2016; Georgiades et al., 2014; Verté et al., 2006), the authors note that subtypes are primarily distinguished in terms of severity of symptoms, rather than that there are qualitative differences between subtypes. This is in contrast with studies that find four subtypes (Hu & Steinberg, 2009; Pichitpunpong et al., 2019), for which there is at least one qualitatively different subtype.

There seems to be some pattern when we look at studies that have applied latent growth curve models: Studies that have used single variables from the Vineland Adaptive Behavior Scales tend to find fewer subtypes (2; Farmer et al., 2018, Bal et al., 2015, Tomaszewski, Smith DaWalt, & Odom, 2019), than studies that have used single variables from the ADOS (4–5, Gotham, Pickles, & Lord, 2012, Venker et al., 2014, Visser et al., 2017). Although the number of subtypes is the same for the VABS studies, the interpretation is different across studies. The initially lower scoring subtypes either increase in score, decrease, or remain stable. For the ADOS studies, the interpretation is more consistent with each study identifying a “severe stable”, a “moderate stable” and a “moderate improving” subtype. The other one to two subtypes differed across studies. We should be careful not to overinterpret these results considering the limited number of studies within each subset, but there seems to be potential in replications by different research groups, as this does give more insight into the robustness of subtyping results.

3.4. Results validation strategies

The prevalence of the various validation strategies is displayed in the left of Fig. 8.

3.4.1. Cross-method replication consists of familiar (parametric or non-parametric) pairs of methods

13% of articles made use of multiple subtyping methods. When this was the case, either multiple non-parametric methods were used, i.e., k-means clustering and hierarchical clustering, or multiple parametric methods were used. The studies that were recorded to perform cross-method replication with parametric methods were often of one of two kinds. The first kind were studies looking at whether interindividual

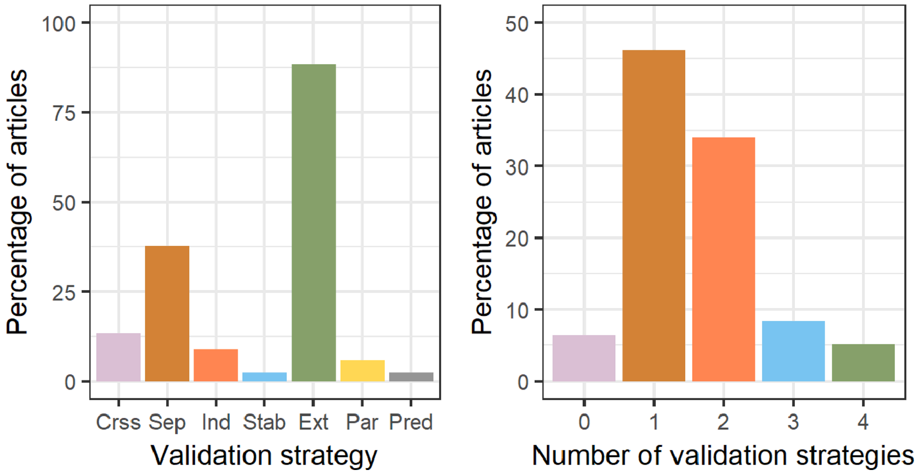

Fig. 8. Percentage of articles that have used the seven validation strategies (left) and the number of validation strategies that were used (right). Note: Crss = Cross-method replication, Sep = Subtype separation, Ind = Independent replication, Stab = Temporal stability, Ext = External validation, Par = Parallel replication, Pred = Predictive validation.

differences are best described with a latent categorical or dimensional structure, for which latent class analyses are compared to factor mixture models and factor models (e.g., Kim et al., 2019; Uljarević et al., 2020). The second kind were studies in which subtypes at one measurement occasion —established with a latent class analysis— are compared to subtypes at a second measurement occasion —with a latent transition analysis. Arguably, this second kind is an example of a temporal stability validation strategy, rather than a cross-method replication. There was only one study that compared results from clustering methods from different traditions, namely Latent Class Analysis and k-means clustering (Uljarević, Frazier, et al., 2020). Apart from this, there were some technical studies that proposed novel methods and compared them to a default method (e.g., Zhao et al., 2018).

3.4.2. *Subtype separation is variable, as different methods come with different metrics and indices*

Subtype separation was investigated in 38% of the articles. The first way of establishing subtype separation we recorded was by computing an index of the difference between subtypes. Used indices included the Calinski-Harabasz index, Dunn index, Davies-Bouldin index, Silhouette index, Gap statistic (Cohen et al., 2017), pseudo-F, pseudo-T^2, and cubic clustering criterion (Ben-Sasson et al., 2008; Lecavalier, 2006). Only some studies go into detail on the meaning of these indices for the validity of the subtypes (e.g., Asif et al., 2020). The second way of establishing subtype separation that was recorded assigned probabilities to people's subtype membership (Ausderau et al., 2014; Voorspoels, Rutten, Bartlema, Tuerlinckx, & Vanpaemel, 2018).

3.4.3. *Independent replication is most commonly observed in samples big enough to split into two*

Independent replication within a single article occurred in 9% of the articles. Some studies had a clear replication design. For example, in a study that made use of data from two schools, the subtyping results were independently replicated, by running the subtyping analysis on each school separately (Cohen et al., 2017). In both schools, two subtypes were identified, that could be interpreted in the same way, i.e., stable sleepers vs. unstable sleepers. The participants in each school were also classified using the subtyping solution from the other school. Other studies featured less direct replications. For example, subtyping results in a child sample were replicated in an adult sample (Lewis, Murdoch, & Woodyatt, 2007b). In this case, if the number of subtypes had not replicated, this could have suggested many things other than that the subtypes were not valid. A number of studies had a sample that was large enough to split into two, establishing an excellent form of independent replication, as the selection of replication participants is made randomly (Lombardo et al., 2016; Uljarević, Frazier, et al., 2020).

3.4.4. *Assessment of temporal stability is rare, although it is recognized as a goal*

Only 3% of studies performed an analysis of temporal stability. As mentioned above in the section on cross-method replication, these were primarily studies that used latent transition analysis to examine stability over time. One study that was particularly explicit in its goals of examining longitudinal stability looked at reading profiles at two measurement occasions, measured 30 months apart (Solari, Grimm, McIntyre, Zajic, & Mundy, 2019). A number of studies explicitly mention investigations of stability as one of the most urgent priorities.

3.4.5. *External validation is common, but authors are rarely explicit about validity implications*

The majority, 88%, of articles describe comparisons between subtypes on variables that were not used to construct the subtypes. By far most often, age was used to compare subtypes. For example, in research in infants, four subtypes that were defined using behaviors scored from a video were found to differ in the age of participants (Malvy et al., 2004). Sex was also frequently used to compare groups. For example, using various self-report measures and tasks measuring empathy to subtype participants, three classes were recovered, one of which was found to be primarily female (Grove, Baillie, Allison, Baron-Cohen, & Hoekstra, 2015). Diagnostic group is a third variable often used to compare subtypes, for example to validate subtypes that are constructed using biological variables (e.g., El-Ansary, Hassan, Daghestani, Al-Ayadhi, & Ben Bacha, 2020). Most articles do not discuss what it means for the validity of the subtypes, whether the subtypes are different on these additional variables or not (notable exceptions in Painter et al., 2018, Vaidya et al., 2020).

3.4.6. *Parallel validation is primarily found in studies with multiple growth curves for multiple variables*

6% of studies performed separate subtyping analyses with similar variables. Two articles used both the Social Responsiveness Scale (SRS), and the Social Communication Questionnaire (SCQ) to form subtypes (same data, Frazier et al., 2010, Frazier et al., 2012). One performed latent growth curve analyses for different measures of daily living skills, with the same subtypes appearing across measures (Bal et al., 2015). One study was unique in that latent transition analysis was not used to model different measurement occasions, but different variables (Spikol, McAteer, & Murphy, 2019); this was coded as parallel validation. The clearest form of parallel validation was found in a study where separate latent growth curves were fitted for four different measures measuring the same construct —symptom onset — three of which were parent-rated, one was examiner-rated (Ozonoff et al., 2018). Interestingly, the analysis indicated different numbers of subtypes for the two types of raters. In one study primarily concerning ADHD symptoms, separate community detection analyses were run for Attention and Hyperactivity measures (Cordova et al., 2020), which was coded as parallel validation, even those these constructs are somewhat different.

3.4.7. *Predictive validation is uncommon*

3% of studies used subtype assignments to make predictions over time. Two studies predicted diagnostic status at age four, using subtypes that were established at age two (Brennan, Barton, Chen, Green, & Fein, 2015; Wiggins, Robins, Adamson, Bakeman, & Henrich, 2012). One study modeled latent growth curves over multiple measurement occasions in early infancy, with which diagnostic status at 36 months was predicted (Nishimura et al., 2019). One study was arguably not predicting but was coded as such, as ASD diagnosis at the last measurement occasion of a latent growth curve model was predicted from subtype membership (Henry et al., 2018).

3.4.8. *Most articles use one or two validation strategies*

In the right of Fig. 8, we display how the frequency of validation strategies is distributed among articles. By far most articles used one or two validation strategies. Use of zero validation strategies mostly occurred in articles that were not trying to make a scientific contribution for ASD per se. For example, some articles use ASD data as an illustration for demonstrating a model-fitting procedure (e.g., Zheng, Hume, Able, Bishop, & Boyd, 2020). There are eight articles that have used four validation strategies (Ausderau et al., 2014; Chen et al., 2019; Chen et al., 2019; Cohen et al., 2017; Duffy & Als, 2019; Obafemi-Ajayi et al., 2015; Solari et al., 2019; Spikol et al., 2019; Uljarević, Frazier, et al., 2020). Ausderau et al. (2014) stands out, as this article is very explicit in the application and reasoning behind using different validators and different validating strategies.

4. Discussion

Much research has been done to establish whether there are subtypes within ASD, and to establish whether ASD can be distinguished from other conditions and typical functioning. This research is highly relevant, as different subtypes may require different interventions, different kinds of care, and may be influenced by different environmental and biological factors. The question is: Is there actually sufficient evidence for the existence of subtypes within ASD? Given the current status of subtyping research we believe that, for many results, there is too little evidence that the observed subtypes are valid and reliable. In general, few of the seven different validation strategies we discussed are applied in the ASD literature. So far, not one single study has been found to apply all seven strategies for validation. This is the case even though our coding of validation strategies was lenient, in the sense that many borderline cases were coded as providing validation. To make the search for subtypes, biomarkers, and tailored interventions truly valuable, it is crucial that researchers a) systematically gather additional variables, independent samples, and follow-up data to validate subtypes, b) preregister hypotheses on what outcomes they expect from these validation strategies, and c) explicitly report what results falsify or confirm the validity of a particular subtyping solution. We are well aware this is not an easy endeavor.

A similar conclusion was reached when Wolfers et al. (2019) focused on a smaller sample of studies, but with the inclusion of pattern classification methods. They particularly stress that more effort should be put towards identifying the biological basis of subtypes. Such a biological basis to subtypes would be one possible approach to link subtypes from multiple domains to each other, if they are found to share the same biological foundation. As described in this article, this need not be the only route to establishing clearly distinct subtypes that can be compared across domains. Also, biological differences do not need to underlie all differences between psychiatric subtypes. On what substrate differences arise depends on the level and domain on which subtyping variables are measured, and on the goal. Furthermore, if the goal is to predict epilepsy, biological factors will be crucial. If the goal is to predict happiness, biological factors will most likely not be sufficient.

Although more inclusion of independent replications within studies would be a great strength (Feczko et al., 2019), it is understandable that many samples within a single study are not large enough to split into two, without sacrificing the statistical power to detect different subtypes. However, between studies, replication of one's own or others' results is possible. In the current sample of studies, there are a number of studies that had such a setup, for example in the sensory studies (Ausderau et al., 2014; Ben-Sasson et al., 2008; Lane, Dennis, & Geraghty, 2011; Lane, Molloy, & Bishop, 2014; Lane, Young, Baker, & Angley, 2010; Uljarević, Lane, Kelly, & Leekam, 2016). This provides the field with the possibility of assessing how replicable the number and composition of subtypes are. By performing a replication using the same measurement instruments and procedure as an existing subtyping study, one may add more to the subtyping literature on ASD than by providing yet another categorization using a novel combination of instruments.

One difficulty for the current state of the validity of subtypes, is that whether a particular result is seen as validation or invalidation is context-dependent. In some studies, correspondence of subtypes with diagnostic categories is seen as a validation of the subtyping result. For example, using gene expression as the subtyping variables, two subtypes were recovered in two studies. The two subtypes were found to correspond with the division into affected and unaffected siblings (Kong et al., 2013) or with the division into the ASD group and control sample (Oh, Kim, Kim, & Ahn, 2017). In contrast, for some other articles, a lack of correspondence between subtype and diagnostic group is seen as an invalidation of diagnostic labels. For example, using subscales of a communication checklist, three subtypes were recovered from a sample of children, which did not correspond one-to-one with the various DSM-IV diagnoses (Autism, Asperger's Syndrome, Pervasive Developmental Disorder – Not Otherwise Specified) that were assigned before the study (Verté et al., 2006). Similarly, using various cognitive measures to make subtypes, four subtypes were recovered that did not correspond one-to-one with the diagnostic labels of ADHD and ASD (Rommelse et al., 2016). These articles use a lack of systematic differences in diagnostic group between subtypes to make a case against the diagnostic labels that are used. It is evident that one can argue both ways. Therefore, it is important that researchers clarify beforehand what result they expect. Preregistration of one's hypotheses and data-analysis plan, through platforms such as AsPredicted.org or the Open Science Framework, are a promising way forward in increasing such transparency (Nosek et al., 2015).

In the discussion of validation, it is beneficial to separate confirmation and falsification of subtyping results. For almost all validation strategies, these require different study parameters or variables. For example, to confirm temporal stability, the researcher may measure the same participants after 10 years on the same variables. If subtypes are identical in type and membership, this provides strong confirming evidence. However, if subtypes are different, this does little for falsifying the earlier found subtypes. The initial result may have overfitted the data, but subtypes may also have changed over time due to developmental processes. To falsify temporal stability, the researcher may measure the same participants twice within a short time frame (weeks to a few months). If subtypes are different, the subtypes are probably too unstable to be useful, which can be counted as a falsification. However, if subtypes remain the same, this provides only weak confirming evidence for their temporal stability. For each strategy, the optimal study design depends on whether the goal is confirmation of subtype validity or subjecting it to possible falsification.

Independent replication can potentially provide the strongest falsification and confirmation which is why we consider this one the most valuable of validation strategies; stronger than for example cross-method replication which provides little opportunity for falsification. Falsification does require that the independent replication is quite direct, as any difference in sampling or diagnostic practice can cause differences in the population from which the researcher is sampling, which in turn can result in true differences in subtypes. The cross-study results on the ADI-R and Vineland should be considered in this context. The convergence of results across Vineland studies provides interesting confirmatory evidence, but the lack of convergence between ADI-R studies does not falsify the subtypes found in any of the studies, as their samples may be representative of —perhaps subtly— different populations.

Some articles are explicit that they do not consider the external variables to provide validation, for example, stating that "classes are descriptively characterized using other phenotypic data" (Farmer et al., 2018). Other researchers make explicit that they consider statistically significant differences on other variables as evidence that the identified subtypes are valid, for example stating that "... comparisons involved an attempt to examine the validity of the clusters" (Brown et al., 2004). or "[t]he validity of the cluster solutions was appraised with data external to the cluster analysis." (Lecavalier, 2006). In the vast majority of the articles, no such reasoning is provided. As mentioned in the results section, most cases of "external validation" were related to descriptions of the subtypes in terms of sex and age.

Although we labelled any comparisons between subtypes on external variables as an attempt at "external validation", for the majority of these articles, we are unsure of the researchers' view on the theoretical implications. Arguably, differences in sex and age neither confirm nor falsify subtypes, even though differences there might suggest subtypes artificially created by the sampling process (e.g., when there are accidental differences between the populations from which older and younger participants are sampled). We would suggest that in the future, researchers are explicit about the theoretical role that external variables play in their analysis, for which a preregistered protocol would again be preferable. One question is whether external validation of subtypes teaches us anything beyond the correlation between subtyping variables and external variables. In other words, are subtypes more than the sum of their parts? We believe so: If we construct subtypes with variables A and B, and external variable C is correlated with neither A nor B, C can still differ between subtypes. Even though this is a theoretical possibility, it would be wise to select candidate external variables that are intermediately correlated with the subtyping variables. External variables that are uncorrelated with the subtyping variables are more likely to be irrelevant, and when external variables are too highly correlated with the subtyping variables, external validation becomes tautological. However, the selection of external variables should be based on a) clinical relevance, b) theoretical plausibility, and c) informativeness for the validity of subtypes, rather than on correlations alone.

Temporal stability is important to research because it matters whether mobility between subtypes is possible, whether there is development, or whether people will always stay in the same subtype: If there is possible mobility between subtypes or development, being part of a subtype with a negative outcome may be a malleable factor. How to calculate temporal stability is difficult to prescribe, even with just two measurement occasions T1 and T2. One could form subtypes and assign separately at T1 and T2, assign at T2 using the subtype specification from T1, explicitly model transitions between T1 and T2, or jointly analyze the data from T1 and T2. All these options require researchers to be explicit about their expectations.

There seems to be a latent and potentially false assumption that the subtypes that are found in some studies will map onto subtypes that are found in other studies. For some part, this may be true, as the most severely affected subtype in one study may well correspond to the most severely affected subtype in another study. However, due to the variety of measures that are used, this is not necessarily the case, and one subtype that is formed on the basis of sensory sensitivities may well be scattered over four different subtypes had the subtyping procedures been based on measures of language abilities. To clarify whether we are referring to the same subtype every time, more studies need to be done that administer multiple types of measures and perform the subtyping analysis for every domain. Then, we can establish whether subtypes are stable across domains, or whether different subtyping solutions are required for different domains. Relatedly, we need to know whether subtypes are stable within a single domain, or whether subtypes are specific to particular measures. This makes evident the need for what we call parallel validation.

Parallel validation is one of the least used forms of validation. This is perhaps because, although psychological and psychiatric theory is formulated on the level of constructs, the bottom-up approach focuses the researcher on subtype differences in the manifest measurement variables. Finding multiple variables that purport to measure the same exact construct is difficult. One strategy could be, if one has sufficient measurements that come from a unidimensional measurement instrument, to randomly select half of the variables, and perform the subtyping analysis on both halves. To our knowledge, such an approach has not been used, but would be valuable to lift discussions up from the level of measurement to the level of theory.

Subtype separation is, after external validation, one of the most used validation methods, but it is still only used in 38% of the studies. However, cluster indices for internal validation of subtypes are becoming increasingly acccessible. Researchers that use Mplus (Muthén & Muthén, 2017) or mclust in R (Scrucca, Fop, Murphy, & Raftery, 2016) are increasingly adding these indices, as they are part of the default output of these software packages. Also, R packages such as NbClust (Charrad, Ghazzali, Boiteau, Niknafs, & Charrad, 2014) offer a variety of indices to be computed and are freely available. Therefore, it seems that the lack of subtype separation may naturally disappear in the future.

When establishing whether a result is validated, it is important to distinguish the different types of similarities in subtyping results that can be achieved. Which one is most important depends on the theoretical background. Ideally, the number of subtypes is the same, subtype sizes are the same, and subtype variable means are the same; regardless of whether one looks at sample 1 or sample 2 in an independent replication, or measurement occasions 1 and 2 in longitudinal stability. There are however many nuances. If mean reaction times for the subtypes differ between measurement occasions, as all participants become older and slower, this is not necessarily an invalidation of longitudinal stability of the subtypes. Also, the relative sizes of subtypes may differ between populations. Therefore, validity is not straightforward, and differences and similarities between subtyping solutions should be interpreted in the light of other evidence.

In this review, we included any study that applied subtyping methods to at least some participants with a diagnosis of ASD. The way ASD is currently conceptualized, the distinctions are not clear-cut between for example ASD and ADHD, and between ASD and some specific personality disorders. Also, people with an ASD diagnosis may on many dimensions have overlapping scores with a non-ASD comparison population. Therefore, to fully capture what the role of the ASD diagnosis is within the hierarchical taxonomy of individual differences, and to discover what is specific to ASD and what is not, one would ideally include studies with other samples as well. This was our reason for also including samples that included other groups, and we included a study where as few as 3% of participants had an ASD diagnosis.

As mentioned in the introduction, developments in the definition of autism could affect the heterogeneity within the population diagnosed with ASD (Mottron & Bzdok, 2020), and by extension, the number of subtypes that would be found. An earlier meta-analysis has shown that, over time, the effects on several domains between groups diagnosed with ASD and comparison groups have been decreasing in size

(Rødgaard, Jensen, Vergnes, Soulières, & Mottron, 2019). This could be due to the ASD diagnosis including more and more people who would not fit the prototypical definitions of autism as used in earlier versions of diagnostic manuals. Although our sample of studies included major shifts in diagnostic manuals —from DSM-IV to DSM-5 the most dramatic— the effect of these shifts on subtypes were not visible in our sample. This is most likely because the other differences between studies already made studies incomparable in this respect. A study with a large population sample to which criteria from multiple versions of diagnostic manuals are applied might be more appropriate to investigate these effects for the subtyping case.

Most important is the practical use of subtypes, which lies in the potential for specific prognoses, estimates of intervention efficiency, and biomarkers. However, predictive validation was among the least used validation strategies. We have only focused on unsupervised learning, i.e., methods that make empirical subtypes from variables, and have excluded supervised approaches, i.e., methods that make predictions from variables. As recently argued, in order to focus subtyping results on having predictive value, unsupervised methods may need to be combined with supervised methods (Feczko et al., 2019). In fact, two of the articles that we have included in this review used random forests, a supervised approach, to establish the similarities between participants, which were then used as input for an unsupervised analysis (Cordova et al., 2020; Feczko et al., 2018). Such combinations of unsupervised and supervised methods potentially form a valuable addition to the subtyping methods currently available, to increase the validity and practical usefulness of subtyping results.

In conclusion, we expect to have clarified where potential improvements lie in the validation of subtyping results when focusing on ASD. However, the same reasoning is also relevant for subtyping in other (clinical) groups. To move the field forward, we need guidelines and recommendations how to validate subtyping results. Below, we provide a subtyping validation checklist. The primary goal of this checklist is to improve the theoretical quality of subtyping results, which also means being clear in what constitutes a validation and an invalidation of a subtyping result. With a systematic approach, we can establish clinically meaningful subtypes that are distinct regardless of statistical method or choice of measurement instruments, replicable, stable over time, and predictive of later difficulties.

4.1. SUbtyping VAlidation Checklist (SUVAC)

To provide guidance for future researchers, we propose a checklist called the SUVAC (for SUbtyping VAlidation Checklist), in Table 1. The SUVAC serves several purposes. The first benefit is that researchers can use the SUVAC in designing their studies, so they can plan for additional variables for parallel validation or external validation, or extra measurement occasions for longitudinal stability. The second benefit is that future systematic reviewers and meta-analysts of subtyping analyses can also use the SUVAC to record different types of validation strategies that have been applied in other fields. The third benefit of the SUVAC is the benefit of common nomenclature. Although most studies used some form of "external validation", a minority of studies called it that explicitly. A lack of common understanding in these terms makes it difficult to evaluate what theoretical conclusions researchers draw from their comparisons. When every study uses the same terminology for "longitudinal stability", the field will more transparent in terms of which subtyping result is stable over time, and which subtyping results are not.

Not all steps are required in every context, and usefulness of subtypes cannot be ensured by following a simple series of steps. This is because validity may be context-dependent. The SUVAC should not be thought of as a checklist of quality —which one can pass or fail— but as a checklist of considerations when planning a subtyping study or evaluating a body of subtyping research. Each of these steps provides a source of evidence for or against the validity and practical usefulness of a subtyping solution. These considerations can provide a foothold to researchers who want to take on the complex task of validating subtypes.

Table 1
SUbtype VAlidation Checklist (SUVAC).

Validation method		+	–	?
Analytical				
Cross-method replication	Are multiple statistical subtyping methods applied in the analysis?			
Subtype separation	Is a standardized metric of distinctiveness of subtypes reported? and/or Is a measure of uncertainty with which participants are assigned to subtypes reported?			
Additional testing of participants				
Independent replication	Is the subtyping analysis repeated in a second sample of participants?			
Temporal stability	Is the subtyping analysis repeated with the same participants at a second measurement occasion?			
Additional variables				
External validation	Are participants from different subtypes compared on variables that were not used in the subtyping analysis?			
Parallel validation	Is the subtyping analysis repeated with a second set of variables, that are purported to measure the same constructs as the variables used in the first subtyping analysis?			
Additional testing of participants + Additional variables				
Predictive validation	Are participants from different subtypes compared on variables that were not used in the subtyping analysis, and that were measured at a second measurement occasion?			
General				
All strategies	Are predictions on the validation steps formulated before the analysis/preregistered?			

Role of funding sources

All authors are supported by a VICI grant awarded by the NWO [grant number 45316-006, awarded to HMG].

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgements

We would like to thank two anonymous reviewers for their constructive comments.

Appendix A. Search terms

(((LATENT adj3 CLASS*) or MIXTURE* or (LATENT adj3 PROFIL*) or (HIDDEN adj2 MARKOV) or (LATENT adj2 MARKOV) or (LATENT adj2

TRAIT) or LATENT VARIABLE or (LATENT adj2 TRAJECTOR*) or (TRAJECTOR* adj2 CLASS) or ((HIERARCH* adj2 CLUSTER*) or (CLUSTER* adj ALGOR*) or (CLUSTER* adj2 ANALY*) or CLUSTERING or K*MEANS or (WARD* adj METHOD) or UNSUPERVISED LEARNING or DATA*DRIVEN) or (COMMUN* DETECT* or CLIQUE*)) and (AUTIS* or ASPERGER* or PERVAS* DEVELOPMENT*)).ti,ab,kw.

Notes: The search terms for PsycINFO and MEDLine were identical, apart from the keyword index (kw in MEDLine, id in PsycINFO).

Appendix B. Table of study characteristics

Table A
Characteristics of the included studies

Article	N, (% ASD diagnosis, Criteria)	% male, mean age (range); mean IQ (range)	Statistical method	Subtyping variables (number)	Validation methods	Number of subtypes (size per subtype in %)
Abu-Akel, Allison, Baron-Cohen, and Heinke (2019)	4717 (17%, Criteria: -)	36% male, Age: 34.47 (18–75), IQ: - (–)	Latent Class Analysis	AQ (1)	Separation	2 (–)
Agelink van Rentergem et al. (2019)	408 (52%, Criteria: DSM-IV)	62% male, Age: 45.3 (19–79), IQ: - (–)	Latent Class Analysis	AQ (50)	External	2 (53, 47)
Al-Jabery et al. (2016)	208 (100%, Criteria: -)	-% male, Age: - (–), IQ: - (–)	Other	ADOS, ADI-R, ABC, PPVT, SRS (11)	Separation, External	2 (91, 9)
Asif et al. (2020)	1397 (100%, Criteria: DSM-IV)	83% male, Age: 7.6 (–), IQ: - (–)	Hierarchical clustering	ADI-R, sex, PIQ, VABS, ADOS (7)	Separation, External	2 (65, 35)
Ausderau et al. (2014)	1294 (100%, Criteria: DSM-IV)	82% male, Age: 7.6 (2–12 at baseline), IQ: 81 (–)	Latent Class Analysis + Latent Transition Analysis	Sensory Experience Questionnaire (4)	Cross-method, Separation, Stability, External	4 (32, 30, 19, 19)
Azad et al. (2020)	476 (65%, Criteria: -)	78% male, Age: 10.2 (5–17), IQ: - (–)	Latent Class Analysis	Pediatric QoL (12)	External	5 (25, 23, 21, 19, 13)
Baez et al. (2020)	287 (36%, Criteria: -)	71% male, Age: 10.2 (8–12), IQ: - (–)	Latent Class Analysis	BRIEF (8)	External	3 (44, 28, 28)
Baeza-Velasco, Michelon, Rattaz, and Baghdadli (2014)	152 (100%, Criteria: ICD-10)	82% male, Age: - (3–7 at baseline), IQ: - (–)	Hierarchical clustering	ABC (4)	Cross-method, Separation, External	4 (36, 34, 18, 13)
Bal et al. (2015)	145 (100%, Criteria: -)	88% male, Age: 2.4 at baseline (–), IQ: 54 (–)	Latent Growth Curve Analysis	VABS (1)	Separation, External, Parallel	2 (66, 34)
Bangerter et al. (2020)	124 (100%, Criteria: -)	75% male, Age: 15 (6–54), IQ: 99 (>60)	Latent Class Analysis	Facial expressions (4)	External	2 (72, 28)
Barrett, Prior, and Manjiviona (2004)	37 (59.5%, Criteria: DSM-IV)	86% male, Age: 5.5 (4–7), IQ: 84 (–)	k-means	Parent-rated social interaction, repetitive behaviors, and pragmatic language, behaviors on video, behavior, IQ (13)	External	3 (45, 31, 24)
Barton et al. (2004)	24 (54%, Criteria: DSM-IV)	71% male, Age: 35 (16–48), IQ: 110 (–)	Hierarchical clustering	Face processing (3)	External	5 (33, 25, 21, 21)
Bathelt et al. (2018)	442 (6%, Criteria: -)	67% male, Age: 9.2 (5–17), IQ: - (–)	Community detection	Conners subscales (6)	Separation, External	3 (34, 33, 33)
Beauchamp et al. (2020)	36 (14%, Criteria: -)	-% male, Age: 7.7 (6–9), IQ: 110 (–)	k-means	Expressive Receptive language (An.1: 2, An.2: 2, An.3: 2, An.4: 4, An.5: 2)	–	An.1: 3, An.2: 3, An.3: 1, An.4: 3, An.5: 2 (An.1: 58, 33, 8, An.2: 56, 36, 8, An.3: 100, An.4: 47, 37, 16, An.5: 63, 37)
Ben-Sasson et al. (2008)	170 (100%, Criteria: -)	78% male, Age: 2.3 (–), IQ: - (–)	Hierarchical clustering	Infant/Toddler Sensory Profile (4)	Cross-method, Separation, External	3 (45, 29, 26)
Berlin et al. (2011)	286 (5%, Criteria: -)	64% male, Age: 3 (–), IQ: - (–)	Latent Class Analysis	Comorbid conditions, Feeding problems (9)	External	3 (58, 37, 5)
Berthoz, Lalanne, Crane, and Hill (2013)	172 (22%, Criteria: DSM-IV-TR)	45% male, Age: 39,1 (–), IQ: - (–)	Latent Class Analysis	AQ, Trait Anxiety, Alexithymia, Anhedonia (10)	–	4 (39, 33, 15, 13)
Bishop-Fitzpatrick et al. (2016)	180 (100%, Criteria: DSM-IV)	75% male, Age: 34 (23–60), IQ: - (–)	Latent Class Analysis	Quality of life outcomes (7)	External	3 (44, 37, 18)
Bitsika et al. (2008)	53 (100%, Criteria: DSM-IV)	81% male, Age: 8.9 (4–12), IQ: 102 (–)	Hierarchical clustering	Age, IQ, VABS, CARS (8)	External	3 (40, 34, 26)
Bitsika, Arnold, and Sharpley (2018)			Two-step cluster analysis	SRS, sensory features, challenging behavior (14)	External	2 (50, 50)

(continued on next page)

(*continued*)

Article	N, (% ASD diagnosis, Criteria)	% male, mean age (range); mean IQ (range)	Statistical method	Subtyping variables (number)	Validation methods	Number of subtypes (size per subtype in %)
	147 (100%, Criteria: DSM-IV-TR)	100% male, Age: 11.21 (6–18), IQ: 95.19 (73–132)				
Brennan et al. (2015)	102 (100%, Criteria: DSM-IV-TR)	76% male, Age: 2.5 (1–2), IQ: - (–)	Hierarchical clustering	ADOS, MSEL (5)	External, Predictive	3 (67, 26, 8)
Bricout et al. (2019)	42 (52%, Criteria: DSM-IV)	100% male, Age: 10.4 (8–12), IQ: - (>70)	Hierarchical clustering	Motor capacities (8)	External	4 (50, 17, 17, 17)
Brown et al. (2004)	308 (8%, Criteria: DSM-IV)	56% male, Age: 13.2 (6–22), IQ: - (ID)	Hierarchical clustering	ABC (4)	Separation, Independent, External	8 (44, 19, 12, 6, 6, 6, 4, 3)
Bureau et al. (2008)	1484 (100%, Criteria: -)	-% male, Age: - (–), IQ: - (–)	Latent Class Analysis	ADI-R (4)	Separation, External	5 (–)
Castro & Pinto, 2015	66 (75.75%, Criteria: DSM-IV-TR/ICD-10)	100% male, Age: 3.2 (2–3), IQ: 67 (–)	k-medians	Inflammatory markers (–)	External	An.1: 2 (56, 44), An.2: 3 (51, 32, 16)
Careaga et al. (2017)	66 (33%, Criteria: -)	-% male, Age: - (3–6), IQ: - (–)	Hierarchical clustering	Matrix for Assessment of Activities and Participation (6)	External	3 (38, 37, 25)
Chen, Uddin, et al., 2019	356 (100%, Criteria: -)	100% male, Age: 14.2 (5–35), IQ: 105.1 (69–148)	k-means	Gray matter volume (60)	Separation, External	3 (53, 29, 18)
Chen, Abrams, et al., 2019	114 (100%, Criteria: -)	100% male, Age: 9.7 (7–12), IQ: 108 (67–150)	Hierarchical clustering + k-means	Numerical Operations, Math Reasoning, Word Reading, Reading Comprehension (4)	Cross-method, Separation, Independent, External	2 (63, 37)
Cholemkery et al. (2016)	463 (100%, Criteria: DSM-5/ICD-10)	88% male, Age: 10.4 (3–21), IQ: 95 (41–147)	Hierarchical clustering	ADI-R (4)	External	3 (37, 30, 33)
Cohen et al. (2017)	106 (100%, Criteria: DSM-IV)	82% male, Age: 14.8 (5–18), IQ: - (<70)	Hierarchical clustering	Sleep features (11)	Separation, Independent, Stability, External	2 (61, 39)
Cordova et al. (2020)	130 (49%, Criteria: DSM-5)	76% male, Age: 11.5 (7–16), IQ: - (–)	Community detection	Executive functioning (43)	External, Parallel	2 (An.1: 61, 39, An.2: 65, 35)
Cuccaro et al. (2012)	577 (100%, Criteria: DSM-IV)	83% male, Age: 9.3 (4–21), IQ: - (>35)	Latent Class Analysis	Age at developmental milestones, ADI-R, VABS (11)	Separation, External	5 (52, 31, 7, 5, 5)
Dajani et al. (2016)	321 (30%, Criteria: DSM-5)	79% male, Age: 10 (8–13), IQ: 109 (63–147)	Latent Class Analysis	BRIEF, NEPSY, Digit Span Backwards (10)	External	3 (43, 33, 24)
DiStefano, Senturk, and Jeste (2019)	33 (100%, Criteria: -)	76% male, Age: 7.6 (5–11), IQ: 63 (–)	k-means	EEG amplitude differences between conditions and latency (3)	Separation, External	4 (48, 33, 12, 6)
Doshi-Velez et al. (2014)	4927 (100%, Criteria: ICD-9)	72% male, Age: 15 (–), IQ: - (–)	Hierarchical clustering	ICD codes (1350)	External	4 (88, 4, 4, 2)
Duffy and Als (2019)	430 (100%, Criteria: DSM-IV-TR/DSM-5)	84% male, Age: 4.7 (2–12), IQ: - (–)	Hierarchical clustering + k-means	EEG coherence factors (40)	Cross-method, Separation, Independent, External	2 (61, 39)
Dyck et al. (2011)	608 (5%, Criteria: -)	55% male, Age: 8.9 (3–14), IQ: - (–)	Latent Class Analysis	IQ, CELF-3, Motor coordination, ToM, Emotion Recognition, TMT, Go/no go (12)	External	2 (–)
Eagle, Romanczyk, and Lenzenweger (2010)	43 (100%, Criteria: DSM-IV-TR)	79% male, Age: 7.1 (2–12), IQ: 62 (40–116)	Latent Class Analysis	IQ, PPVT, Autism behavior inventory, Social interaction inventory (4)	Separation, External	2 (63, 37)
Easson, Fatima, and McIntosh (2019)	266 (55%, Criteria: -)	100% male, Age: 16.3 (6–39), IQ: 108.7 (76–148)	k-means	Correlations between ROI pairs (–)	Separation, External	2 (52, 48)
El-Ansary et al. (2020)	37 (35%, Criteria: -)	-% male, Age: - (2–14), IQ: - (–)	Hierarchical clustering	Biomarkers (An.1: 9, An.2: 5, An.3: 14)	External	2 (65, 35)
Elwin, Schröder, Ek, Wallsten, and Kjellin (2017)	71 (100%, Criteria: ICD-10)	37% male, Age: - (18–65), IQ: - (ID excluded)	Hierarchical clustering	Sensory Reactivity in Autism Spectrum (4)	External	3 (52, 24, 24)
Farmer et al. (2018)	105 (100%, Criteria: DSM-IV-TR)	88% male, Age: 4.3 at baseline (1–7), IQ: 50 (–)	Latent Growth Curve Analysis	VABS (1)	Separation, External	2 (73, 27)
Feczko et al. (2018)	47 (100%, Criteria: DSM-IV)	77% male, Age: 12.15 (9–13), IQ: - (–)	Community detection	Information processing tasks (34)	Separation, External	3 (53, 28, 19)
Fountain, Winter, and Bearman (2012)	6975 (100%, Criteria: DSM-IV-TR)	82% male, Age: - (–), IQ: - (–)	Latent Growth Curve Analysis	Client Development Evaluation Report (3)	Separation, External	Soc: 6 (27, 25, 19, 13, 11, 7), Com: 6 (30, 25,

(*continued on next page*)

(*continued*)

Article	N, (% ASD diagnosis, Criteria)	% male, mean age (range); mean IQ (range)	Statistical method	Subtyping variables (number)	Validation methods	Number of subtypes (size per subtype in %)
						20, 10, 8, 7), RRB: 6 (29, 27, 23, 8, 7, 6)
Frazier et al. (2010)	An. 1: 11472, An. 2 & 3: 4400 (60%, Criteria: DSM-IV-TR)	68% male, Age: 8.2 (–), IQ: - (–)	Latent Class Analysis	An.1: SCQ, An.2: SRS, An.3: SRS (3)	Parallel	2 (–)
Frazier et al. (2012)	6949 (61%, Criteria: DSM-IV-TR)	68% male, Age: 8.4 (4–18), IQ: - (–)	Factor Mixture Model	SRS (8)	External, Parallel	2 (63, 37)
Garon et al. (2009)	34 (100%, Criteria: DSM-IV-TR)	65% male, Age: 2 at baseline (–), IQ: 81 (–)	Two-step cluster analysis	Sex, Age of Dx, IQ, ADOS, TBAQ-R (7)	–	2 (53, 47)
Georgiades, Szatmari, Boyle, Hanna, et al. (2013)	391 (100%, Criteria: DSM-IV)	84% male, Age: 3.2 (2–5), IQ: - (–)	Factor Mixture Model	ADI-R (26)	External	3 (56, 34, 10)
Georgiades et al. (2014)	280 (100%, Criteria: DSM-IV)	86% male, Age: 3.4 at baseline (2–4), IQ: - (–)	Factor Mixture Model	ADI-R (26)	Stability, External	3 at baseline, 2 at retest (55, 35, 9)
Gizzonio, Avanzini, Fabbri-Destro, Campi, and Rizzolatti (2014)	95 (33%, Criteria: DSM-IV-TR)	53% male, Age: 8.7 (6–16), IQ: 101 (–)	k-means	Wechsler Intelligence Scale for Children (10)	External	3, fixed (–)
Gonthier, Longuépée, and Bouvard (2016)	148 (100%, Criteria: DSM-IV-TR)	70% male, Age: 33 (19–59), IQ: - (ID)	Hierarchical clustering + k-means	Adult Sensory Profile (4)	Cross-method, External	4 (30, 30, 24, 16)
Gotham et al. (2012)	345 (97%, Criteria: DSM-IV-TR)	82% male, Age: 3.3 at baseline (2–15), IQ: 61 (–)	Latent Growth Curve Analysis	ADOS (1)	Separation, External	4 (46, 38, 9, 7)
Greaves-Lord et al. (2013)	949 (100%, Criteria: DSM-IV-TR)	82% male, Age: 9.3 (6–18), IQ: - (–)	Latent Class Analysis	CBCL & Children's Social Behavior Questionnaire (14)	Separation, External	6 (30, 23, 15, 12, 12, 8)
Grove et al. (2015)	1034 (35%, Criteria: -)	44% male, Age: 38 (–), IQ: - (–)	Latent Class Analysis + Factor Mixture Model	EQ, SQ, AQ, RMET, Emotional Faces (6)	Cross-method, Separation	3 in FMM, 4 in LCA (45, 30, 25)
Harper-Hill, Copland, and Arnott (2013)	35 (57%, Criteria: -)	74% male, Age: 11.4 (9–16), IQ: 100 (–)	Hierarchical clustering	CELF-4, CNRep (2)	External	2 (83, 17)
Hasenstab et al. (2016)	37 (100%, Criteria: -)	-% male, Age: 4.5 (2–6), IQ: 78 (49–123)	Other	ERP (–)	Separation	2, fixed (71, 29)
Henry et al. (2018)	91 (13%, Criteria: DSM-5)	62% male, Age: 1.6 (1), IQ: - (–)	Latent Growth Curve Analysis	MSEL (2)	External, Parallel, Predictive	Verb.: 3, Non-verb.: 2 (Verb.: 66, 23, 11, Non-verb.: 82, 18)
Hoogenhout and Malcolm-Smith (2017)	62 (100%, Criteria: DSM-IV)	85% male, Age: 11.2 (8–16), IQ: 81 (50–123)	Hierarchical clustering	Theory of Mind (11)	Separation, External	3 (48, 42, 10)
Hrdlicka et al. (2005)	64 (100%, Criteria: ICD-10)	81% male, Age: 9.4 (3–15), IQ: - (ID)	Hierarchical clustering	MRI (11)	External	4 (52, 28, 14, 6)
Hu and Steinberg (2009)	1954 (100%, Criteria: -)	78% male, Age: 8.3 (1–48), IQ: - (–)	Hierarchical clustering + k-means + Figure of Merit	ADI-R (123)	Cross-method	4 (–)
Ingalhalikar et al. (2012)	54 (61%, Criteria: -)	-% male, Age: - (–), IQ: - (–)	Other	Regions of Interest, IQ, SRS, SCQ, ADOS, PRI (85)	–	2 (–)
Jao Keehn et al. (2019)	57 (100%, Criteria: DSM-5)	82% male, Age: 13.8 (9–18), IQ: 104.4 (66–141)	k-means	ROI pairs (3)	Separation, External	2 (56, 44)
Ji et al. (2011)	293 (39%, Criteria: DSM-IV)	76% male, Age: 7.6 (2–21), IQ: 37 (–)	Hierarchical clustering	ABC (4)	External	4 (39, 24, 20, 18)
Kamp-Becker et al. (2010)	140 (74%, Criteria: DSM-IV/ICD-10)	94% male, Age: 12.3 (6–24), IQ: 93 (70–139)	Hierarchical clustering	An.1: ADI-R (13), An.2: ADOS (8), An.3: Cognitive functioning (8), An.4: ADI-R (–), An.5: ADOS (28)	External	An.1: 1 (100), An.2: 3 (97, –, –), An.3: 2 (99), An.4: 3 (77, 19, 4), An.5: 2 (78, 22)
Kang, Gadow, and Lerner (2020)	223 (100%, Criteria: DSM-IV)	80% male, Age: 10.5 (6–18), IQ: 86.1 (–)	Latent Class Analysis	Atypical communication characteristics (13)	External	4 (34, 31, 30, 4)
Katsuki, Yamashita, Yamane, Kanba, and Yoshida (2020)	314 (59%, Criteria: DSM-5)	82% male, Age: 8.9 (4–15), IQ: 95.4 (>70)	Hierarchical clustering	CBCL (8)	Separation, External	4 (37, 29, 17, 17)
Kim et al. (2019)	3825 (68%, Criteria: DSM-IV-TR)	73% male, Age: 11.35 (6–22), IQ: - (–)	Latent Class Analysis + Factor Mixture Model	Child and Adolescent Symptom Inventory-4R ASD subscale (12)	Cross-method, Independent, External	An.1: 6, An.2: 2, An.3. 6 (–)
Kim and Ha (2019)	333 (28%, Criteria: DSM-5)	77% male, Age: 2.7 (1–5), IQ: - (–)	Hierarchical clustering	CBCL (7)	External	3 (43, 30, 28)
Klopper et al. (2017)			Hierarchical clustering	ADI-R, ADOS (88)		2 (64, 36)

(*continued on next page*)

(*continued*)

Article	N, (% ASD diagnosis, Criteria)	% male, mean age (range); mean IQ (range)	Statistical method	Subtyping variables (number)	Validation methods	Number of subtypes (size per subtype in %)
	61 (100%, Criteria: DSM-IV-TR/DSM-5)	84% male, Age: 8.8 (5–14), IQ: 106 (>70)			Separation, External	
Kong et al. (2013)	40 (50%, Criteria: SSC)	62% male, Age: 10.6 (4–17), IQ: - (–)	Hierarchical clustering	Differentially expressed probesets (189)	External	2 (53, 48)
Kushki et al. (2019)	226 (50%, Criteria: -)	75% male, Age: 11.3 (6–18), IQ: 101.3 (–)	Other	Cortical regions + autism, inattention, obsessive compulsion (79)	Separation, Independent, External	10 (13, 13, 13, 10, 10, 9, 9, 8, 8, 6)
Kyriakopoulos et al. (2015)	84 (100%, Criteria: ICD-10)	75% male, Age: 11.1 (–), IQ: - (–)	Latent Class Analysis	Symptoms in case notes (8)	External	2 (51, 49)
LaBianca et al. (2018)	55 (38%, Criteria: ICD-10)	49% male, Age: 33.6 (–), IQ: - (–)	Hierarchical clustering	Demographic, clinical, functional characteristics (47)	Separation, External	3 (60, 31, 9)
Landa et al. (2012)	204 (25%, Criteria: -)	46% male, Age: - (0–1 at baseline), IQ: - (–)	Multivariate Latent Growth Curve Model	MSEL (5)	External	4 (40, 26, 22, 12)
Lane et al. (2010)	54 (100%, Criteria: DSM-IV-TR)	87% male, Age: 6.6 (2–9), IQ: - (–)	Latent Class Analysis	SSP (8)	External	3 (44, 32, 24,)
Lane et al. (2011)	29 (100%, Criteria: -)	80% male, Age: 6.7 (3–9), IQ: - (–)	Latent Class Analysis	SSP (7)	External	5 (31, 21, 21, 14, 14)
Lane et al. (2014)	228 (100%, Criteria: DSM-IV-TR)	89% male, Age: 5.1 (2−10), IQ: 72 (21−132)	Latent Class Analysis	SSP (7)	External	4 (40, 38, 12, 10)
Lecavalier (2006)	Parent: 353, Teacher: 437 (Parent: 65, Teacher: 72%, Criteria: -)	83% male, Age: 9.6 (3–21), IQ: - (–)	Hierarchical clustering + k-means	Nisonger Child Behavior Rating Form (8)	Cross-method, Separation, External	Parent: 6 (31, 21, 14, 13, 13, 9), Teacher: 8 (23, 20, 13, 12, 11, 8, 8, 5)
Lerner, De Los Reyes, Drabick, Gerber, and Gadow (2017)	218 (100%, Criteria: DSM-IV)	82% male, Age: 10.5 (6–18), IQ: 86 (–)	Latent Class Analysis	CASI (6)	Separation, External	4 (44, 29, 17, 10)
Lewis, Murdoch, and Woodyatt (2007a)	Child: 20, Adult: 17 (100%, Criteria: DSM-IV)	65% male, Age: Children: 11.5, Adults: 35 (Children: 9–17, Adults: 18–67), IQ: - (–)	Hierarchical clustering	Test of Language Competence (4)	Independent, External	Child: 3 (55, 25, 20), Adult: 3 (47, 29, 24)
Lewis et al. (2007b)	20 (100%, Criteria: DSM-IV)	80% male, Age: 11.6 (9–17), IQ: - (–)	Hierarchical clustering	CELF-4 (5)	External	4 (35, 30, 20, 15)
Lewis et al. (2008)	17 (100%, Criteria: DSM-IV)	47% male, Age: 35 (18–67), IQ: 91 (–)	Hierarchical clustering	Right Hemisphere Language Battery, Western Aphasia Battery (7)	External	2 (59, 41)
Lindly, Chan, Levy, Parker, and Kuhlthau (2020)	1378 (100%, Criteria: DSM-IV-TR/DSM-5)	85% male, Age: 10.7 (6–18), IQ: 82.1 (–)	Latent Class Analysis	Service use (13)	Separation, External	4 (43, 36, 12, 9)
Lingren et al. (2016)	20,658 (100%, Criteria: ICD-9)	-% male, Age: - (–), IQ: - (–)	k-means	ICD codes (100−200)	Separation, Independent	Coh.1: 5 (80, 8, 5, 4, 3); Coh.2: 5 (62, 14, 10, 8, 7); Coh.3: 6 (46, 15, 14, 14, 8, 4)
Liss, Saulnier, Fein, and Kinsbourne (2006)	144 (100%, Criteria: DSM-IV)	80% male, Age: 8.5 (–), IQ: - (–)	Hierarchical clustering	VABS, Sensory Questionnaire, Interview (12)	Cross-method, External	4 (33, 31, 25, 12)
Little et al. (2017)	1132 (9%, Criteria: -)	57% male, Age: 8.2 (–), IQ: - (–)	Latent Class Analysis	Child Sensory Profile 2 (86)	Separation, External	5 (78, 7, 5, 5, 4)
Lombardo et al. (2016)	694 (56.91%, Criteria: DSM-5/ICD-10)	48% male, Age: - (18–74), IQ: - (–)	Hierarchical clustering	RMET (36)	Independent, External	9 (–)
Malvy et al. (2004)	74 (100%, Criteria: DSM-IV/ICD-10)	57% male, Age: 1.8 (0–2), IQ: 56 (25–105)	Hierarchical clustering	Behaviors on video (4)	External	4 (36, 24, 23, 16)
Matta, Zhao, Ercal, and Obafemi-Ajayi (2018)	2680 (100%, Criteria: -)	86% male, Age: - (4–17), IQ: 78.5 (–)	Other	ADOS, ADI, VABS, SRS, RBS, CBCL, IQ (36)	External	5 (–)
McChesney and Toseeb (2018)	13,210 (3%, Criteria: -)	50% male, Age: 11 (11−11), IQ: - (–)	Latent Class Analysis	Happiness, Self-esteem, Prosociality (3)	External	5 (61, 23, 7, 6, 3)
McCrimmon et al. (2012)	66 (50%, Criteria: DSM-IV-TR)	79% male, Age: 19 (16–21), IQ: 112 (At least 85)	Two-step cluster analysis	D-KEFS (7)	External	2 (55, 45)
McIntyre et al. (2017)	81 (100%, Criteria: -)	82% male, Age: 11.2 (8–16), IQ: 100 (76–132)	Latent Class Analysis	Reading ability (12)	External	4 (33, 32, 20, 14)
			Hierarchical clustering		External	3 (48, 38, 13)

(*continued on next page*)

(*continued*)

Article	N, (% ASD diagnosis, Criteria)	% male, mean age (range); mean IQ (range)	Statistical method	Subtyping variables (number)	Validation methods	Number of subtypes (size per subtype in %)
Mira, Berenguer, Roselló, Baixauli, and Miranda (2019)	52 (100%, Criteria: DSM-5)	92% male, Age: - (7–11), IQ: 101 (>80)		Age, ADHD + ASD symptoms, daily living skills, emotional and behavioral problems, pragmatic competence, ToM (7)		
Montgomery et al. (2018)	188 (100%, Criteria: DSM-IV-TR)	84% male, Age: 10 (3–27), IQ: 80 (20–140)	Latent Class Analysis	NVIQ, ADOS, ADI-R, VABS (7)	External	3 (60, 25, 15)
Morris et al. (2016)	531 (–%, Criteria: DSM-5)	46% male, Age: 11 (2–83), IQ: 86 (56–117)	Factor Mixture Model	SRS (13)	External	2 (81, 19)
Mulder et al. (2004)	77 (100%, Criteria: DSM-IV-TR)	86% male, Age: 12.5 (–), IQ: - (–)	Latent Class Analysis	Platelet serotonin (1)	–	2 (–)
Munson et al. (2008)	456 (100%, Criteria: -)	81% male, Age: 3.6 (2–5), IQ: - (–)	Latent Class Analysis	MSEL (4)	Separation, External	4 (59, 22, 13, 7)
Nevill, Hedley, Uljarević, Butter, and Mulick (2017)	158 (77%, Criteria: DSM-5)	87% male, Age: 2.7 (1–3), IQ: - (–)	Hierarchical clustering + k-means	VABS (4)	Cross-method, External	2 (69, 31)
Nishimura et al. (2019)	952 (3%, Criteria: DSM-IV)	51% male, Age: 2 (at end) (0–2), IQ: - (–)	Multivariate Latent Growth Curve Model	MSEL (5)	External	5 (49, 21, 14, 12, 4)
Nordahl et al. (2020)	300 (100%, Criteria: -)	70% male, Age: 3 (2–5), IQ: - (–)	Latent Class Analysis	CBCL, VABS, ADOS, DQ (10)	Separation, External	3 (40, 32, 27)
Obafemi-Ajayi et al. (2015)	62 (100%, Criteria: DSM-IV)	100% male, Age: - (8–12), IQ: 85 (31–130)	Expectation Maximization + k-means + Self-Organizing feature Map + Partioning Around Medoids	Facial features (31)	Cross-method, Separation, Independent, External	3 (48, 29, 23)
Obara et al. (2018)	17 (100%, Criteria: -)	76% male, Age: 9 (5–19), IQ: - (–)	Affinity propagation, k-medoids	Concomitant clumsiness and sound sensitivity, plasma glutamine, autism rating scale (3)	Cross-method, External	5 (35, 24, 18, 12, 12)
Oh et al. (2017)	42 (50%, Criteria: DSM-IV-TR)	81% male, Age: 27 (–), IQ: 92 (–)	Hierarchical clustering	Differentially expressed probe-sets (19)	External	2 (57, 43)
Ozonoff et al. (2011)	52 (100%, Criteria: DSM-IV)	88% male, Age: 3.3 at baseline (1–4), IQ: - (–)	Latent Growth Curve Analysis	Behaviors on video (1)	Separation, External	3 (38, 38, 23)
Ozonoff et al. (2018)	32 (100%, Criteria: DSM-5)	66% male, Age: 0.3 at baseline (–), IQ: - (–)	Latent Growth Curve Analysis	Symptom onset (4)	Separation, Parallel	Examiner: 2, Parent: 3 (Examiner: 88, 12, Parent: 69, 19, 12)
Painter et al. (2018)	1692 (4%, Criteria: -)	55% male, Age: 42 (18–90), IQ: ID sample (ID)	Hierarchical clustering + k-means	Learning Disability Needs Assessment Tool (23)	Cross-method, External	6 (28, 19, 18, 13, 11, 10)
Parikh, Kurzius-Spencer, Mastergeorge, and Pettygrove (2018)	2303 (100%, Criteria: DSM-IV-TR)	20% male, Age: 8 (8), IQ: - (–)	Latent Class Analysis	Race, special education category, delay in social, delay in language, regression of skills (5)	Separation, External	5 (33, 28, 22, 13, 4)
Paynter, Trembath, and Lane (2018)	210 (100%, Criteria: -)	80% male, Age: 4 (2–5), IQ: - (–)	Latent Class Analysis	Change scores in MSEL and VABS (8)	External	2 (55, 45)
Pichitpunpong et al. (2019)	85 (100%, Criteria: -)	100% male, Age: 12 (5–28), IQ: - (–)	Hierarchical clustering	ADI-R (123)	External	4 (35, 28, 24, 13)
Pickles et al. (2014)	192 (84%, Criteria: DSM-IV)	84% male, Age: 2.4 at baseline (–), IQ: 67 (–)	Multivariate Latent Growth Curve Model	Expressive Receptive language (2)	Separation, External	7 (32, 23, 19, 13, 5, 4, 3)
Pickles, McCauley, Pepa, Huerta, and Lord (2020)	123 (90%, Criteria: -)	85% male, Age: 26 (–), IQ: 64 (3–133)	Latent Class Analysis	ADOS, Work, Living, Friends, Number of medicines, PANAS, ADI-R, Adult Behavior Checklist, ABC, VABS (15)	Separation, External	4 (27, 26, 25, 22)
Pohl et al. (2014)	830 (50%, Criteria: -)	0% male, Age: 38 (–), IQ: - (–)	Latent Class Analysis	Sex-steroid linked symptoms (11)	External	2, fixed (64, 36)
Pry, Bodet, Pernon, Aussilloux, and Baghdadli (2007)	207 (100%, Criteria: ICD-10)	81% male, Age: 5.4 (2–7 at baseline), IQ: - (–)	Hierarchical clustering	Object-related cognition, person-related cognition, VABS (5)	External	4 (29, 25, 24, 21)
Qian (2018)	112 (100%, Criteria: -)	88% male, Age: 4 (2–5), IQ: - (–)	k-means	MSEL & Preschool Language Scale (3)	External	3 (38, 36, 27)
Rapin et al. (2009)	62 (100%, Criteria: DSM-III)	86% male, Age: 8.6 (6–9), IQ: 89 (–)	Hierarchical clustering	Expressive phonology, Language comprehension (2)	External	4 (65, 18, 11, 6)
Ring, Woodbury-Smith, Watson, Wheelwright, and	333 (100%, Criteria: DSM-IV)	67% male, Age: - (16–78), IQ: - (–)	Hierarchical clustering	AQ (50)	–	4 (42, 36, 12, 10)

(*continued on next page*)

(*continued*)

Article	N, (% ASD diagnosis, Criteria)	% male, mean age (range); mean IQ (range)	Statistical method	Subtyping variables (number)	Validation methods	Number of subtypes (size per subtype in %)
Baron-Cohen (2008)						
Rommelse et al. (2016)	254 (–%, Criteria: DSM-IV)	64% male, Age: 11.3 (5–17), IQ: 105 (>70)	Latent Class Analysis	Cognitive measures (9)	External	4 (39, 24, 20, 17)
Ros and Graziano (2019)	100 (37%, Criteria: -)	75% male, Age: 4.7 (–), IQ: - (>65)	Latent Class Analysis	Executive functioning, Emotion regulation (6)	External, Predictive	4 (36, 25, 22, 17)
Rubenstein et al. (2019)	707 (100%, Criteria: DSM-5)	82% male, Age: 4.9 (2–5.7), IQ: - (–)	Latent Class Analysis	ADI-R, CBCL, MSEL, ADOS, SCQ, Early Development, Gastrointestinal (25)	External	4 (33, 28, 27, 12.)
Sacco et al. (2012)	245 (100%, Criteria: DSM-IV)	88% male, Age: 8.8 (2–30), IQ: - (–)	Hierarchical clustering + k-means	Circadian and sensory dysfunction, Immune dysfunction, Neurodevelopmental delay, Stereotypic behavior (4)	Cross-method, External	4 (34, 31, 18, 18)
Seynhaeve and Nader-Grosbois (2008)	24 (50%, Criteria: DSM-IV)	79% male, Age: 3.6 (–), IQ: 46 (–)	Hierarchical clustering	Dysregulation disorder, Social and cognitive development (An.1: 6, An.2: 3)	External	An.1: 2 (54, 46), An.2: 2 (54, 46)
Shen et al. (2007)	358 (100%, Criteria: DSM-IV)	86% male, Age: 6.9 (2–21), IQ: - (–)	Expectation Maximization + k-means + Hierarchical clustering	ADI-R (22)	Cross-method, Separation, External	4 (39, 32, 15, 14)
Shogren et al. (2017)	1062 (100%, Criteria: -)	76% male, Age: 10.3 (5–16), IQ: - (ID)	Latent Class Analysis	Support need questionnaire (7)	External	4 (–)
Silleresi et al. (2020)	43 (100%, Criteria: DSM-5/ ICD-11)	96% male, Age: 8.9 (6–12), IQ: 92 (69–125)	k-means	Factor scores of language, non-verbal IQ, autism severity (2)	–	5 (An. 1: 26, 26, 26, 12, 12. An. 2: 33, 26, 21, 12, 9)
Simpson, Adams, Alston-Knox, Heussler, and Keen (2019)	248 (100%, Criteria: -)	82% male, Age: 7.6 (4–11), IQ: - (–)	Dirichlet process mixture	SSP (4)	Separation, External	2 (73, 27)
Smith, Mirenda, and Zaidman-Zait (2007)	35 (100%, Criteria: DSM-IV)	80% male, Age: 3.8 at baseline (1–5 at baseline), IQ: - (48–63)	Hierarchical clustering	Communicative Development Inventory (1)	External	4 (43, 23, 20, 14)
Solari et al. (2019)	80 (100%, Criteria: -)	81% male, Age: 11.3 (7–15), IQ: 100 (–)	Latent Class Analysis + Latent Transition Analysis	Language (12)	Cross-method, Separation, Stability, External	4 (34, 24, 24, 18)
Solomon et al. (2018)	102 (100%, Criteria: -)	80% male, Age: - (1–3 at baseline), IQ: - (–)	Latent Class Analysis	MSEL (1)	Separation, External	4 (35, 26, 22, 18)
Spiker, Lotspeich, Dimiceli, Myers, and Risch (2002)	351 (100%, Criteria: DSM-IV)	79% male, Age: 9.3 (–), IQ: 65 (–)	k-means	ADI-R, verbal status, non-verbal IQ (12)	–	3 (55, 38, 7)
Spikol et al. (2019)	7977 (–%, Criteria: ICD-10)	52% male, Age: 10.5 (4–17), IQ: - (–)	Latent Class Analysis + Latent Transition Analysis	Autism questions on a survey (An. 1: 5, An. 2: 10)	Cross-method, Separation, External, Parallel	3 (An. 1: 87, 11, 2. An. 2: 52, 32, 17)
Stevens et al. (2019)	2400 (100%, Criteria: DSM-5)	81% male, Age: 7.8 (2–12), IQ: - (–)	Latent Class Analysis	Skill domains (8)	External	GMM: 16, Hierarchical clustering: 5 (GMM: -. Hierarchical clustering: 35, 32, 14, 12, 6)
Storlie et al. (2018)	486 (–%, Criteria: -)	-% male, Age: - (–), IQ: - (–)	Dirichlet process mixture	Behavior checklists, ADOS, Woodcock, SRS, SCQ, etc. (55)	–	3 (–)
Sullivan, Gallagher, and Heron (2019)	2079 (100%, Criteria: -)	86% male, Age: 10.3 (5–18), IQ: 81.5 (7–167)	Latent Class Analysis	IQ, VABS, ADOS, CBCL, Aberrant Behavior Checklist, RBS (12)	Separation	5 (25, 24, 20, 20, 12)
Tanaka et al. (2017)	113 (100%, Criteria: DSM-IV-TR)	76% male, Age: 8.2 (3–12), IQ: 105 (80–146)	Hierarchical clustering	CCC-2 (10)	External	2 (78, 22)
Tomaszewski et al. (2019)	244 (100%, Criteria: -)	85% male, Age: 16.4 (13–20), IQ: 84.6 (–)	Latent Growth Curve Analysis	VABS scales (3)	External, Parallel	Comm.: 2, Daily Living: 2, Soc.: 2 (Comm: 87, 13. Daily Living: 89, 11. Soc. 82, 18)
Tomchek, Little, Myers, and Dunn (2018)	400 (100%, Criteria: DSM-IV-TR)	87% male, Age: 4.1 (3–6), IQ: - (–)	Latent Class Analysis	SSP and adaptive, social, language, motor skills (12)	Separation, External	4 (51, 25, 15, 10)
Trantou, Carlsen, Anderson, and Steingrimsson (2021)	516 (22%, Criteria: -)	51% male, Age: 23.9 (19–29), IQ: - (–)	Latent Class Analysis	Sex, age, comorbid diagnosis with ADHD, sickness absence (3)	External	An. 1: 3, An. 2: 4 (An. 1: 54, 34, 12. An. 2: 43, 34, 19, 3)
Uljarević et al. (2016)	57 (100%, Criteria: -)	-% male, Age: 14.2 (11–17), IQ: - (–)	Latent Class Analysis	SSP (7)	External	3 (50, 33, 16)

(*continued on next page*)

(*continued*)

Article	N, (% ASD diagnosis, Criteria)	% male, mean age (range); mean IQ (range)	Statistical method	Subtyping variables (number)	Validation methods	Number of subtypes (size per subtype in %)
Uljarević, Frazier, et al. (2020)	13,282 (60%, Criteria: -)	69% male, Age: 8.1 (2–18), IQ: - (–)	Latent Class Analysis + Factor Mixture Model	SCQ (19)	Cross-method, Separation, Independent, External	3 (–)
Uljarević et al. (2020)	164 (100%, Criteria: -)	79% male, Age: 7.5 (3–17), IQ: 75.2 (14–122)	Latent Class Analysis + k-means clustering	Social abilities (5)	Cross-method, Separation, External	5 (–)
Vaidya et al. (2020)	320 (30%, Criteria: DSM-IV-TR,DSM-5)	63% male, Age: 10.6 (8–14), IQ: 110.8 (>70)	Community detection	ADHD-RS, BRIEF, CBCL (12)	Separation, Independent, External	3 (43, 30, 26)
van der Meer et al. (2012)	644 (–%, Criteria: DSM-IV)	56% male, Age: 9.9 (5–17), IQ: 105 (>70)	Latent Class Analysis	SCQ, CPRS (13)	External	5 (42, 23, 17, 9, 9)
Vargason, Frye, McGuinness, and Hahn (2019)	3278 (100%, Criteria: ICD-9)	82% male, Age: - (0–10), IQ: - (–)	k-means	Medical records (70)	External	3 (50, 27, 24)
Veatch, Veenstra-VanderWeele, Potter, Pericak-Vance, and Haines (2014)	1261 (100%, Criteria: DSM-IV)	80% male, Age: - (2–21), IQ: - (–)	Hierarchical clustering	Age, VABS, ADI-R, ADOS, Head circumference (13)	Separation, Independent, External	An.1: 2 (65, 35), An.2: 10
Venker et al. (2014)	129 (100%, Criteria: DSM-IV)	87% male, Age: - (2–3 at baseline), IQ: 76 (38–115)	Latent Growth Curve Analysis	ADOS (1)	External	4 (42, 36, 14, 8)
Verté et al. (2006)	135 (100%, Criteria: DSM-IV-TR)	90% male, Age: 8.8 (6–13), IQ: 101 (>80)	Hierarchical clustering	ADI-R (12)	External	3 (47, 34, 19)
Visser et al. (2017)	203 (–%, Criteria: DSM-IV)	82% male, Age: 2.7 at baseline (1–4), IQ: 71 (–)	Latent Growth Curve Analysis	ADOS (1)	External	5 (48, 22, 20, 5, 5)
Voorspoels et al. (2018)	40 (50%, Criteria: -)	90% male, Age: - (7–12), IQ: 110 (–)	Latent Class Analysis	Stimuli categorization (60)	Separation, External	2, fixed (78, 20, 1 person not classifiable)
Waddington et al. (2018)	675 (23%, Criteria: -)	55% male, Age: 12.6 (7–18), IQ: 103 (>70)	Latent Class Analysis + Factor Mixture Model	Emotion recognition (4)	Cross-method, Separation, External	4 (45, 26, 22, 7)
Walker, Langefeld, Zimmerman, Schwartz, and Krigsman (2019)	35 (100%, Criteria: DSM-IV)	83% male, Age: 7.6 (3–17), IQ: - (–)	Hierarchical clustering	Differentially expressed transcripts (68)	External	2 (54, 46)
Wei, Christiano, Yu, Wagner, and Spiker (2015)	130 (100%, Criteria: -)	86% male, Age: 7.6 (6–9), IQ: - (–)	k-means	Reading and math achievement (5)	Independent, External	4 (39, 32, 20, 9)
Wiggins et al. (2012)	186 (100%, Criteria: DSM-IV)	80% male, Age: 2.2 at baseline (1–3), IQ: 61 (49–127)	Hierarchical clustering	CARS (15)	External, Predictive	3 (51, 25, 24)
Wiggins et al. (2017)	707 (100%, Criteria: DSM-5)	82% male, Age: 4.9 (2–5), IQ: - (–)	Latent Class Analysis	ADI-R, CBCL, MSEL, ADOS, SCQ, Early Development, Gastrointestinal (25)	External	4 (34, 28, 26, 12)
Wiggins et al. (2019)	672 (100%, Criteria: -)	82% male, Age: - (2–5), IQ: - (–)	Latent Class Analysis	Abilities, skills, age at developmental milestones, etc. (19)	Separation, External	4 (34, 28, 26, 12)
Zaidman-Zait et al. (2020)	178 (100%, Criteria: DSM-IV)	88% male, Age: 10.6 (10–11), IQ: 76.4 (–)	Latent Class Analysis	WIAT, Teacher Report Form, VABS, CCC (6)	Separation, External	4 (31, 24, 24, 21)
Zheng et al. (2020)	188 (100%, Criteria: -)	84% male, Age: 4 (2–5), IQ: 70 (9–130)	Hierarchical clustering	Abilities, behaviors and skills (9)	–	3 (51, 25, 25)

Notes: - means missing, ABC:Aberrant Behavior Checklist, ADI-R:Autism Diagnostic Interview-Revised, ADOS:Autism Diagnostic Observation Schedule, AQ: Autism spectrum Quotient, ASD: Autism Spectrum Disorder, BDI:Beck Depression Inventory, BRIEF:Behavior Rating Inventory of Executive Function, CARS:Child Autism Rating Scale, CASI:Child and Adolescent Symptom Inventory, CBCL:Child Behavior Checklist, CCC:Children's Communication Checklist, CELF: Clinical Evaluation of Language Fundamentals, CNrep:Children's Non-Word Repetition, CPRS:Children's Psychiatric Rating Scale, D-KEFS:Delis-Kaplan Executive Function System, DSM: Diagnostic and Statistical Manual of Mental Disorders, Dx:Diagnosis, EQ:Empathy Quotient, ERP:Event-related potential, FMM:Factor Mixture Model, ICD:International Classification of Diseases, ID:Intellectual Disability, IQ:Intelligence Quotient, LCA:Latent Class Analysis, MRI:Magnetic Resonance Imaging, MSEL:Mullen Scales of Early Learning, NEPSY:Developmental Neuropsychological Assessment, PIQ: Performance IQ, PPVT:Peabody Picture Vocabulary, PRI:Perceptual Reasoning Index, RBS:Repetitive Behavior Scale, RMET:Reading the Mind in the Eyes Test, SCQ:Social Communication Questionnaire, SQ:Systemising Quotient Revised, SRS:Social Responsiveness Scale, SSC:Simons Simplex Collection, SSP: Short Sensory Profile, TBAQ-R: Toddler Behavior Assessment Questionnaire-Revised, TMT:Trail Making Test, ToM:Theory of Mind, VABS:Vineland Adaptive Behavior Scales, WIAT = Wechsler Individual Achievement Test.

References

Abu-Akel, A., Allison, C., Baron-Cohen, S., & Heinke, D. (2019). The distribution of autistic traits across the autism spectrum: Evidence for discontinuous dimensional subpopulations underlying the autism continuum. *Molecular Autism, 10*(1), 24.

Agelink van Rentergem, J. A., Lever, A. G., & Geurts, H. M. (2019). Negatively phrased items of the autism spectrum quotient function differently for groups with and without autism. *Autism, 23*(7), 1752–1764.

Al-Jabery, K., Obafemi-Ajayi, T., Olbricht, G. R., Takahashi, T. N., Kanne, S., & Wunsch, D. (2016). Ensemble statistical and subspace clustering model for analysis of autism spectrum disorder phenotypes. In *38th annual international conference of the IEEE engineering in medicine and biology society* (pp. 3329–3333). IEEE.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Asif, M., Martiniano, H. F., Marques, A. R., Santos, J. X., Vilela, J., Rasga, C., … Vicente, A. M. (2020). Identification of biological mechanisms underlying a multidimensional ASD phenotype using machine learning. *Translational Psychiatry, 10*(43), 1–12.

Ausderau, K. K., Furlong, M., Sideris, J., Bulluck, J., Little, L. M., Watson, L. R., … Baranek, G. T. (2014). Sensory subtypes in children with autism spectrum disorder: Latent profile transition analysis using a national survey of sensory features. *Journal of Child Psychology and Psychiatry, 55*(8), 935–944.

Azad, G. F., Dillon, E., Feuerstein, J., Kalb, L., Neely, J., & Landa, R. (2020). Quality of life in school-aged youth referred to an autism specialty clinic: A latent profile analysis. *Journal of Autism and Developmental Disorders*, 1–12.

Baez, A. C., Dajani, D. R., Voorhies, W., Parladé, M. V., Alessandri, M., Britton, J. C., … Uddin, L. Q. (2020). Parsing heterogeneity of executive function in typically and atypically developing children: A conceptual replication and exploration of social function. *Journal of Autism and Developmental Disorders, 50*(3), 707–718.

Baeza-Velasco, C., Michelon, C., Rattaz, C., & Baghdadli, A. (2014). Are aberrant behavioral patterns associated with the adaptive behavior trajectories of teenagers with autism spectrum disorders? *Research in Autism Spectrum Disorders, 8*(3), 304–311.

Bal, V. H., Kim, S.-H., Cheong, D., & Lord, C. (2015). Daily living skills in individuals with autism spectrum disorder from 2 to 21 years of age. *Autism, 19*(7), 774–784.

Bangerter, A., Chatterjee, M., Manfredonia, J., Manyakov, N. V., Ness, S., Boice, M. A., … Leventhal, B. (2020). Automated recognition of spontaneous facial expression in individuals with autism spectrum disorder: Parsing response variability. *Molecular Autism, 11*(1), 1–15.

Barrett, S., Prior, M., & Manjiviona, J. (2004). Children on the borderlands of autism: Differential characteristics in social, imaginative, communicative and repetitive behavior domains. *Autism, 8*(1), 61–87.

Barton, J. J., Cherkasova, M. V., Hefter, R., Cox, T. A., O'connor, M., & Manoach, D. S. (2004). Are patients with social developmental disorders prosopagnosic? Perceptual heterogeneity in the asperger and socio-emotional processing disorders. *Brain, 127* (8), 1706–1716.

Bathelt, J., Holmes, J., Astle, D. E., Gathercole, S., Astle, D., Manly, T., & Kievit, R. (2018). Data-driven subtyping of executive function–related behavioral problems in children. *Journal of the American Academy of Child & Adolescent Psychiatry, 57*(4), 252–262.

Beauchamp, M. L., Rezzonico, S., & MacLeod, A. A. (2020). Bilingualism in school-aged children with ASD: A pilot study. *Journal of Autism and Developmental Disorders*, 1–16.

Becerra, T. A., von Ehrenstein, O. S., Heck, J. E., Olsen, J., Arah, O. A., Jeste, S. S., … Ritz, B. (2014). Autism spectrum disorders and race, ethnicity, and nativity: A population-based study. *Pediatrics, 134*(1), Article e63.

Beglinger, L. J., & Smith, T. H. (2001). A review of subtyping in autism and proposed dimensional classification model. *Journal of Autism and Developmental Disorders, 31* (4), 411–422.

Ben-Sasson, A., Cermak, S., Orsmond, G., Tager-Flusberg, H., Kadlec, M., & Carter, A. (2008). Sensory clusters of toddlers with autism spectrum disorders: Differences in affective symptoms. *Journal of Child Psychology and Psychiatry, 49*(8), 817–825.

Berlin, K. S., Lobato, D. J., Pinkos, B., Cerezo, C. S., & LeLeiko, N. S. (2011). Patterns of medical and developmental comorbidities among children presenting with feeding problems: A latent class analysis. *Journal of Developmental & Behavioral Pediatrics, 32* (1), 41–47.

Bernstein, A., Stickle, T. R., Zvolensky, M. J., Taylor, S., Abramowitz, J., & Stewart, S. (2010). Dimensional, categorical, or dimensional-categories: Testing the latent structure of anxiety sensitivity among adults using factor-mixture modeling. *Behavior Therapy, 41*(4), 515–529.

Bernstein, A., Zvolensky, M. J., Norton, P. J., Schmidt, N. B., Taylor, S., Forsyth, J. P., … Cox, B. (2007). Taxometric and factor analytic models of anxiety sensitivity: Integrating approaches to latent structural research. *Psychological Assessment, 19*(1), 74.

Berthoz, S., Lalanne, C., Crane, L., & Hill, E. L. (2013). Investigating emotional impairments in adults with autism spectrum disorders and the broader autism phenotype. *Psychiatry Research, 208*(3), 257–264.

Bishop-Fitzpatrick, L., Hong, J., Smith, L. E., Makuch, R. A., Greenberg, J. S., & Mailick, M. R. (2016). Characterizing objective quality of life and normative outcomes in adults with autism spectrum disorder: An exploratory latent class analysis. *Journal of Autism and Developmental Disorders, 46*(8), 2707–2719.

Bitsika, V., Arnold, W. M., & Sharpley, C. F. (2018). Cluster analysis of autism spectrum disorder symptomatology: Qualitatively distinct subtypes or quantitative degrees of severity of a single disorder? *Research in Developmental Disabilities, 76*, 65–75.

Bitsika, V., Sharpley, C., & Orapeleng, S. (2008). An exploratory analysis of the use of cognitive, adaptive and behavioural indices for cluster analysis of ASD subgroups. *Journal of Intellectual Disability Research, 52*(11), 973–985.

Bohane, L., Maguire, N., & Richardson, T. (2017). Resilients, overcontrollers and undercontrollers: A systematic review of the utility of a personality typology method in understanding adult mental health problems. *Clinical Psychology Review, 57*, 75–92.

Brennan, L., Barton, M., Chen, C.-M., Green, J., & Fein, D. (2015). Detecting subgroups in children diagnosed with pervasive developmental disorder-not otherwise specified. *Journal of Autism and Developmental Disorders, 45*(5), 1329–1344.

Brewin, C. R., Cloitre, M., Hyland, P., Shevlin, M., Maercker, A., Bryant, R. A., … Somasundaram, D. (2017). A review of current evidence regarding the ICD-11 proposals for diagnosing PTSD and complex PTSD. *Clinical Psychology Review, 58*, 1–15.

Bricout, V. A., Pace, M., Dumortier, L., Miganeh, S., Mahistre, Y., & Guinot, M. (2019). Motor capacities in boys with high functioning autism: Which evaluations to choose? *Journal of Clinical Medicine, 8*(10), 1521.

Brown, E. C., Aman, M. G., & Lecavalier, L. (2004). Empirical classification of behavioral and psychiatric problems in children and adolescents with mental retardation. *American Journal on Mental Retardation, 109*(6), 445–455.

Brugha, T. S., McManus, S., Bankart, J., Scott, F., Purdon, S., Smith, J., … Meltzer, H. (2011). Epidemiology of autism spectrum disorders in adults in the community in England. *Archives of General Psychiatry, 68*(5), 459–465.

Bureau, A., Labbe, A., Croteau, J., & Mérette, C. (2008). Using disease symptoms to improve detection of linkage under genetic heterogeneity. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 32*(5), 476–486.

Careaga, M., Rogers, S., Hansen, R. L., Amaral, D. G., Van de Water, J., & Ashwood, P. (2017). Immune endophenotypes in children with autism spectrum disorder. *Biological Psychiatry, 81*(5), 434–441.

Castro, S., & Pinto, A. (2015). Matrix for assessment of activities and participation: Measuring functioning beyond diagnosis in young children with disabilities. *Developmental Neurorehabilitation, 18*(3), 177–189.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package "nbclust". *Journal of Statistical Software, 61*, 1–36.

Chen, H., Uddin, L. Q., Guo, X., Wang, J., Wang, R., Wang, X., … Chen, H. (2019). Parsing brain structural heterogeneity in males with autism spectrum disorder reveals distinct clinical subtypes. *Human Brain Mapping, 40*(2), 628–637.

Chen, L., Abrams, D. A., Rosenberg-Lee, M., Iuculano, T., Wakeman, H. N., Prathap, S., … Menon, V. (2019). Quantitative analysis of heterogeneity in academic achievement of children with autism. *Clinical Psychological Science, 7*(2), 362–380.

Cholemkery, H., Medda, J., Lempp, T., & Freitag, C. M. (2016). Classifying autism spectrum disorders by ADI-R: Subtypes or severity gradient? *Journal of Autism and Developmental Disorders, 46*(7), 2327–2339.

Cohen, S., Fulcher, B. D., Rajaratnam, S. M., Conduit, R., Sullivan, J. P., St Hilaire, M. A., … Ahearn, W. (2017). Behaviorally-determined sleep phenotypes are robustly associated with adaptive functioning in individuals with low functioning autism. *Scientific Reports, 7*(1), 1–8.

Cordova, M., Shada, K., Demeter, D. V., Doyle, O., Miranda-Dominguez, O., Perrone, A., … Nigg, J. (2020). Heterogeneity of executive function revealed by a functional random forest approach across ADHD and ASD. *NeuroImage: Clinical*, 102245.

Cuccaro, M. L., Tuchman, R. F., Hamilton, K. L., Wright, H. H., Abramson, R. K., Haines, J. L., … Pericak-Vance, M. (2012). Exploring the relationship between autism spectrum disorder and epilepsy using latent class cluster analysis. *Journal of Autism and Developmental Disorders, 42*(8), 1630–1641.

Dajani, D. R., Llabre, M. M., Nebel, M. B., Mostofsky, S. H., & Uddin, L. Q. (2016). Heterogeneity of executive functions among comorbid neurodevelopmental disorders. *Scientific Reports, 6*, 36566.

DeBoth, K. K., & Reynolds, S. (2017). A systematic review of sensory-based autism subtypes. *Research in Autism Spectrum Disorders, 36*, 44–56.

DiStefano, C., Senturk, D., & Jeste, S. S. (2019). ERP evidence of semantic processing in children with ASD. *Developmental Cognitive Neuroscience, 36*, 100,640.

Doshi-Velez, F., Ge, Y., & Kohane, I. (2014). Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics, 133*(1), Article e54.

Duffy, F. H., & Als, H. (2019). Autism, spectrum or clusters? An EEG coherence study. *BMC Neurology, 19*(1), 27.

Dyck, M. J., Piek, J. P., & Patrick, J. (2011). The validity of psychiatric diagnoses: The case of "specific"developmental disorders. *Research in Developmental Disabilities, 32* (6), 2704–2713.

Dziak, J. J., Lanza, S. T., & Tan, X. (2014). Effect size, statistical power, and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 534–552.

Eack, S. M., Bahorik, A. L., McKnight, S. A., Hogarty, S. S., Greenwald, D. P., Newhill, C. E., … Minshew, N. J. (2013). Commonalities in social and non-social cognitive impairments in adults with autism spectrum disorder and schizophrenia. *Schizophrenia Research, 148*(1–3), 24–28.

Eagle, R. F., Romanczyk, R. G., & Lenzenweger, M. F. (2010). Classification of children with autism spectrum disorders: A finite mixture modeling approach to heterogeneity. *Research in Autism Spectrum Disorders, 4*(4), 772–781.

Easson, A. K., Fatima, Z., & McIntosh, A. R. (2019). Functional connectivity-based subtypes of individuals with and without autism spectrum disorder. *Network Neuroscience, 3*(2), 344–362.

El-Ansary, A., Hassan, W. M., Daghestani, M., Al-Ayadhi, L., & Ben Bacha, A. (2020). Preliminary evaluation of a novel nine-biomarker profile for the prediction of autism spectrum disorder. *PLoS One, 15*(1), Article e0227626.

Elsabbagh, M., Divan, G., Koh, Y. J., Kim, Y. S., Kauchali, S., Marcín, C., … Yasamy, M. T. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism Research, 5*(3), 160–179.

Elwin, M., Schröder, A., Ek, L., Wallsten, T., & Kjellin, L. (2017). Sensory clusters of adults with and without autism spectrum conditions. *Journal of Autism and Developmental Disorders, 47*(3), 579–589.

Farmer, C., Swineford, L., Swedo, S. E., & Thurm, A. (2018). Classifying and characterizing the development of adaptive behavior in a naturalistic longitudinal study of young children with autism. *Journal of Neurodevelopmental Disorders, 10*(1), 1.

Feczko, E., Balba, N. M., Miranda-Dominguez, O., Cordova, M., Karalunas, S. L., Irwin, L., … Van Santen, J. (2018). Subtyping cognitive profiles in autism spectrum disorder using a functional random forest algorithm. *Neuroimage, 172*, 674–688.

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The heterogeneity problem: Approaches to identify psychiatric subtypes. *Trends in cognitive sciences, 23*(7), 584–601.

Fombonne, E. (2018). The rising prevalence of autism. *Journal of Child Psychology and Psychiatry, 59*(7), 717–720.

Fountain, C., Winter, A. S., & Bearman, P. S. (2012). Six developmental trajectories characterize children with autism. *Pediatrics, 129*(5), Article e1112.

Frazier, T. W., Youngstrom, E. A., Sinclair, L., Kubu, C. S., Law, P., Rezai, A., … Eng, C. (2010). Autism spectrum disorders as a qualitatively distinct category from typical behavior in a large, clinically ascertained sample. *Assessment, 17*(3), 308–320.

Frazier, T. W., Youngstrom, E. A., Speer, L., Embacher, R., Law, P., Constantino, J., … Eng, C. (2012). Validation of proposed dsm-5 criteria for autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(1), 28–40.

Garon, N., Bryson, S. E., Zwaigenbaum, L., Smith, I. M., Brian, J., Roberts, W., & Szatmari, P. (2009). Temperament and its relationship to autistic symptoms in a high-risk infant sib cohort. *Journal of Abnormal Child Psychology, 37*(1), 59–78.

Georgiades, S., Boyle, M., Szatmari, P., Hanna, S., Duku, E., Zwaigenbaum, L., … Smith, I. (2014). Modeling the phenotypic architecture of autism symptoms from time of diagnosis to age 6. *Journal of Autism and Developmental Disorders, 44*(12), 3045–3055.

Georgiades, S., Szatmari, P., & Boyle, M. (2013). Importance of studying heterogeneity in autism. *Neuropsychiatry, 3*(2), 123.

Georgiades, S., Szatmari, P., Boyle, M., Hanna, S., Duku, E., Zwaigenbaum, L., … Smith, I. (2013). Investigating phenotypic heterogeneity in children with autism spectrum disorder: A factor mixture modeling approach. *Journal of Child Psychology and Psychiatry, 54*(2), 206–215.

Gizzonio, V., Avanzini, P., Fabbri-Destro, M., Campi, C., & Rizzolatti, G. (2014). Cognitive abilities in siblings of children with autism spectrum disorders. *Experimental Brain Research, 232*(7), 2381–2390.

Gonthier, C., Longuépée, L., & Bouvard, M. (2016). Sensory processing in low-functioning adults with autism spectrum disorder: Distinct sensory profiles and their relationships with behavioral dysfunction. *Journal of Autism and Developmental Disorders, 46*(9), 3078–3089.

Gotham, K., Pickles, A., & Lord, C. (2012). Trajectories of autism severity in children using standardized ados scores. *Pediatrics, 130*(5), Article e1278.

Greaves-Lord, K., Eussen, M. L., Verhulst, F. C., Minderaa, R. B., Mandy, W., Hudziak, J. J., … Hartman, C. A. (2013). Empirically based phenotypic profiles of children with pervasive developmental disorders: Interpretation in the light of the DSM-5. *Journal of Autism and Developmental Disorders, 43*(8), 1784–1797.

Grove, R., Baillie, A., Allison, C., Baron-Cohen, S., & Hoekstra, R. A. (2015). Exploring the quantitative nature of empathy, systemising and autistic traits using factor mixture modelling. *The British Journal of Psychiatry, 207*(5), 400–406.

Grzadzinski, R., Huerta, M., & Lord, C. (2013). DSM-5 and autism spectrum disorders (asds): An opportunity for identifying asd subtypes. *Molecular Autism, 4*(1), 12.

Happé, F., Ronald, A., & Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience, 9*(10), 1218.

Harper-Hill, K., Copland, D., & Arnott, W. (2013). Do spoken nonword and sentence repetition tasks discriminate language impairment in children with an ASD? *Research in Autism Spectrum Disorders, 7*(2), 265–275.

Hasenstab, K., Sugar, C., Telesca, D., Jeste, S., & Şentürk, D. (2016). Robust functional clustering of ERP data with application to a study of implicit learning in autism. *Biostatistics, 17*(3), 484–498.

Henry, L., Farmer, C., Manwaring, S. S., Swineford, L., & Thurm, A. (2018). Trajectories of cognitive development in toddlers with language delays. *Research in Developmental Disabilities, 81*, 65–72.

Hoogenhout, M., & Malcolm-Smith, S. (2017). Theory of mind predicts severity level in autism. *Autism, 21*(2), 242–252.

Hrdlicka, M., Dudova, I., Beranova, I., Lisy, J., Belsan, T., Neuwirth, J., … Blatny, M. (2005). Subtypes of autism by cluster analysis based on structural MRI data. *European Child & Adolescent Psychiatry, 14*(3), 138–144.

Hu, V. W., & Steinberg, M. E. (2009). Novel clustering of items from the autism diagnostic interview-revised to define phenotypes within autism spectrum disorders. *Autism Research, 2*(2), 67–77.

Ingalhalikar, M., Smith, A. R., Bloy, L., Gur, R., Roberts, T. P., & Verma, R. (2012). Identifying sub-populations via unsupervised cluster analysis on multi-edge similarity graphs. In *International conference on medical image computing and computer-assisted intervention* (pp. 254–261).

Jao Keehn, R. J., Nair, S., Pueschel, E. B., Linke, A. C., Fishman, I., & Müller, R. A. (2019). Atypical local and distal patterns of occipito-frontal functional connectivity are related to symptom severity in autism. *Cerebral Cortex, 29*(8), 3319–3330.

Ji, N. Y., Capone, G. T., & Kaufmann, W. (2011). Autism spectrum disorder in down syndrome: Cluster analysis of aberrant behaviour checklist data supports diagnosis. *Journal of Intellectual Disability Research, 55*(11), 1064–1077.

Kamp-Becker, I., Smidt, J., Ghahreman, M., Heinzel-Gutenbrunner, M., Becker, K., & Remschmidt, H. (2010). Categorical and dimensional structure of autism spectrum disorders: The nosologic validity of asperger syndrome. *Journal of Autism and Developmental Disorders, 40*(8), 921–929.

Kang, E., Gadow, K. D., & Lerner, M. D. (2020). Atypical communication characteristics, differential diagnosis, and the autism spectrum disorder phenotype in youth. *Journal of Clinical Child & Adolescent Psychology, 49*(2), 251–263.

Kanner, L. (1943). Autistic disturbances of affective contact. *The Nervous Child, 2*(3), 217–250.

Katsuki, D., Yamashita, H., Yamane, K., Kanba, S., & Yoshida, K. (2020). Clinical subtypes in children with attention-deficit hyperactivity disorder according to their child behavior checklist profile. *Child Psychiatry & Human Development*, 1–9.

Kim, H., Keifer, C., Rodriguez-Seijas, C., Eaton, N., Lerner, M., & Gadow, K. (2019). Quantifying the optimal structure of the autism phenotype: A comprehensive comparison of dimensional, categorical, and hybrid models. *Journal of the American Academy of Child & Adolescent Psychiatry, 58*(9), 876–886.

Kim, J. Y., & Ha, E. H. (2019). Cluster analysis of the child behavior checklist 1.5–5 for preschool children diagnosed with a mental disorder. *Psychological Reports, 00*, 1–22.

Klopper, F., Testa, R., Pantelis, C., & Skafidas, E. (2017). A cluster analysis exploration of autism spectrum disorder subgroups in children without intellectual disability. *Research in Autism Spectrum Disorders, 36*, 66–78.

Kong, S., Shimizu-Motohashi, Y., Campbell, M., Lee, I., Collins, C., Brewster, S., … Kunkel, L. (2013). Peripheral blood gene expression signature differentiates children with autism from unaffected siblings. *Neurogenetics, 14*(2), 143–152.

Kushki, A., Anagnostou, E., Hammill, C., Duez, P., Brian, J., Iaboni, A., … Lerch, J. P. (2019). Examining overlap and homogeneity in ASD, ADHD, and OCD: A data-driven, diagnosis-agnostic approach. *Translational Psychiatry, 9*(1), 1–11.

Kyriakopoulos, M., Stringaris, A., Manolesou, S., Radobuljac, M. D., Jacobs, B., Reichenberg, A., … Frangou, S. (2015). Determination of psychosis-related clinical profiles in children with autism spectrum disorders using latent class analysis. *European Child & Adolescent Psychiatry, 24*(3), 301–307.

LaBianca, S., Pagsberg, A. K., Jakobsen, K. D., Demur, A. B., Bartalan, M., LaBianca, J., & Werge, T. (2018). Clusters and trajectories across the autism and/or ADHD spectrum. *Journal of Autism and Developmental Disorders, 48*(10), 3629–3636.

Lai, M.-C., Lombardo, M. V., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/gender differences and autism: Setting the scene for future research. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*(1), 11–24.

Landa, R. J., Gross, A. L., Stuart, E. A., & Bauman, M. (2012). Latent class analysis of early developmental trajectory in baby siblings of children with autism. *Journal of Child Psychology and Psychiatry, 53*(9), 986–996.

Lane, A. E., Dennis, S. J., & Geraghty, M. E. (2011). Brief report: Further evidence of sensory subtypes in autism. *Journal of Autism and Developmental Disorders, 41*(6), 826–831.

Lane, A. E., Molloy, C. A., & Bishop, S. L. (2014). Classification of children with autism spectrum disorder by sensory subtype: A case for sensory-based phenotypes. *Autism Research, 7*(3), 322–333.

Lane, A. E., Young, R. L., Baker, A. E., & Angley, M. T. (2010). Sensory processing subtypes in autism: Association with adaptive behavior. *Journal of Autism and Developmental Disorders, 40*(1), 112–122.

Lecavalier, L. (2006). Behavioral and emotional problems in young people with pervasive developmental disorders: Relative prevalence, effects of subject characteristics, and empirical classification. *Journal of Autism and Developmental Disorders, 36*(8), 1101–1114.

Lerner, M. D., De Los Reyes, A., Drabick, D. A., Gerber, A. H., & Gadow, K. D. (2017). Informant discrepancy defines discrete, clinically useful autism spectrum disorder subgroups. *Journal of Child Psychology and Psychiatry, 58*(7), 829–839.

Lewis, F. M., Murdoch, B. E., & Woodyatt, G. C. (2007a). Communicative competence and metalinguistic ability: Performance by children and adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 37*(8), 1525–1538.

Lewis, F. M., Murdoch, B. E., & Woodyatt, G. C. (2007b). Linguistic abilities in children with autism spectrum disorder. *Research in Autism Spectrum Disorders, 1*(1), 85–100.

Lewis, F. M., Woodyatt, G. C., & Murdoch, B. E. (2008). Linguistic and pragmatic language skills in adults with autism spectrum disorder: A pilot study. *Research in Autism Spectrum Disorders, 2*(1), 176–187.

Lindly, O. J., Chan, J., Levy, S. E., Parker, R. A., & Kuhlthau, K. A. (2020). Service use classes among school-aged children from the autism treatment network registry. *Pediatrics, 145*(s1), 140–150.

Lingren, T., Chen, P., Bochenek, J., Doshi-Velez, F., Manning-Courtney, P., Bickel, J., … Barbaresi, W. (2016). Electronic health record based algorithm to identify patients with autism spectrum disorder. *PLoS One, 11*(7). e0159621.

Liss, M., Saulnier, C., Fein, D., & Kinsbourne, M. (2006). Sensory and attention abnormalities in autistic spectrum disorders. *Autism, 10*(2), 155–172.

Little, L., Dean, E., Tomchek, S., & Dunn, W. (2017). Classifying sensory profiles of children in the general population. *Child: Care, Health and Development, 43*(1), 81–88.

Lombardo, M. V., Lai, M. C., Auyeung, B., Holt, R. J., Allison, C., Smith, P., … Bailey, A. J. (2016). Unsupervised data-driven stratification of mentalizing heterogeneity in autism. *Scientific Reports, 6*(1), 1–15.

Malvy, J., Barthélémy, C., Damie, D., Lenoir, P., Bodier, C., & Roux, S. (2004). Behaviour profiles in a population of infants later diagnosed as having autistic disorder. *European Child & Adolescent Psychiatry, 13*(2), 115–122.

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., & Beckmann, C. F. (2016). Beyond lumping and splitting: A review of computational approaches for stratifying psychiatric disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1*(5), 433–447.

Matta, J., Zhao, J., Ercal, G., & Obafemi-Ajayi, T. (2018). Applications of node-based resilience graph theoretic framework to clustering autism spectrum disorders phenotypes. *Applied Network Science, 3*(1), 38.

McChesney, G., & Toseeb, U. (2018). Happiness, self-esteem, and prosociality in children with and without autism spectrum disorder: Evidence from a UK population cohort study. *Autism Research, 11*(7), 1011–1023.

McCrimmon, A. W., Schwean, V. L., Saklofske, D. H., Montgomery, J. M., & Brady, D. I. (2012). Executive functions in asperger's syndrome: An empirical investigation of verbal and nonverbal skills. *Research in Autism Spectrum Disorders, 6*(1), 224–233.

McIntyre, N. S., Solari, E. J., Grimm, R. P., Lerro, L. E., Gonzales, J. E., & Mundy, P. C. (2017). A comprehensive examination of reading heterogeneity in students with high functioning autism: Distinct reading profiles and their relation to autism symptom severity. *Journal of Autism and Developmental Disorders, 47*(4), 1086–1101.

McKay, D., Abramowitz, J. S., Calamari, J. E., Kyrios, M., Radomsky, A., Sookman, D., … Wilhelm, S. (2004). A critical evaluation of obsessive–compulsive disorder subtypes: Symptoms versus mechanisms. *Clinical Psychology Review, 24*(3), 283–313.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50*(4), 266.

van der Meer, J. M., Oerlemans, A. M., van Steijn, D. J., Lappenschaar, M. G., de Sonneville, L. M., Buitelaar, J. K., & Rommelse, N. N. (2012). Are autism spectrum disorder and attention-deficit/hyperactivity disorder different manifestations of one overarching disorder? Cognitive and symptom evidence from a clinical and population-based sample. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(11), 1160–1172.

Mira, Á., Berenguer, C., Roselló, B., Baixauli, I., & Miranda, A. (2019). Exploring the profiles of children with autism spectrum disorder: Association with family factors. *International Journal of Developmental Disabilities*, 1–11.

Montgomery, A. K., Shuffrey, L. C., Guter, S. J., Anderson, G. M., Jacob, S., Mosconi, M. W., … Veenstra-VanderWeele, J. (2018). Maternal serotonin levels are associated with cognitive ability and core symptoms in autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 57*(11), 867–875.

Morris, S. M., Acosta, M. T., Garg, S., Green, J., Huson, S., Legius, E., … Weiss, L. A. (2016). Disease burden and symptom structure of autism in neurofibromatosis type 1: A study of the international NF1-ASD consortium team (INFACT). *JAMA Psychiatry, 73*(12), 1276–1284.

Mottron, L., & Bzdok, D. (2020). Autism spectrum heterogeneity: Fact or artifact? *Molecular Psychiatry*, 1–8.

Mulder, E. J., Anderson, G. M., Kema, I. P., De Bildt, A., Van Lang, N. D., Den Boer, J. A., & Minderaa, R. B. (2004). Platelet serotonin levels in pervasive developmental disorders and mental retardation: Diagnostic group differences, within-group distribution, and behavioral correlates. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(4), 491–499.

Munson, J., Dawson, G., Sterling, L., Beauchaine, T., Zhou, A., Koehler, E., … Abbott, R. (2008). Evidence for latent classes of IQ in young children with autism spectrum disorder. *American Journal on Mental Retardation, 113*(6), 439–452.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (2005). *Group-based modeling of development*. Harvard University Press.

Nevill, R. E., Hedley, D., Uljarević, M., Butter, E., & Mulick, J. A. (2017). Adaptive behavior profiles in young children with autism spectrum disorder diagnosed under dsm-5 criteria. *Research in Autism Spectrum Disorders, 43*, 53–66.

Nishimura, T., Takei, N., & Tsuchiya, K. J. (2019). Neurodevelopmental trajectory during infancy and diagnosis of autism spectrum disorder as an outcome at 32 months of age. *Epidemiology, 30*, S9–S14.

Nordahl, C. W., Iosif, A. M., Young, G. S., Hechtman, A., Heath, B., Lee, J. K., … Ozonoff, S. (2020). High psychopathology subgroup in young children with autism: associations with biological sex and amygdala volume. *Journal of the American Academy of Child & Adolescent Psychiatry, 59*(12), 1353–1363.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Contestabile, M. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425.

Obafemi-Ajayi, T., Miles, J. H., Takahashi, T. N., Qi, W., Aldridge, K., Zhang, M., … Duan, Y. (2015). Facial structure analysis separates autism spectrum disorders into meaningful clinical subgroups. *Journal of Autism and Developmental Disorders, 45*(5), 1302–1317.

Obara, T., Ishikuro, M., Tamiya, G., Ueki, M., Yamanaka, C., Mizuno, S., … Kobayashi, T. (2018). Potential identification of vitamin B6 responsiveness in autism spectrum disorder utilizing phenotype variables and machine learning methods. *Scientific Reports, 8*(1), 1–7.

Oh, D. H., Kim, I. B., Kim, S. H., & Ahn, D. H. (2017). Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning. *Clinical Psychopharmacology and Neuroscience, 15*(1), 47.

Ozonoff, S., Gangi, D., Hanzel, E. P., Hill, A., Hill, M. M., Miller, M., … Iosif, A. M. (2018). Onset patterns in autism: Variation across informants, methods, and timing. *Autism Research, 11*(5), 788–797.

Ozonoff, S., Iosif, A. M., Young, G. S., Hepburn, S., Thompson, M., Colombi, C., … Rogers, S. J. (2011). Onset patterns in autism: correspondence between home video and parent report. *Journal of the American Academy of Child & Adolescent Psychiatry, 50*(8), 796–806.

Painter, J., Ingham, B., Trevithick, L., Hastings, R. P., & Roy, A. (2018). Identifying needs-based groupings among people accessing intellectual disability services. *American Journal on Intellectual and Developmental Disabilities, 123*(5), 426–442.

Parikh, C., Kurzius-Spencer, M., Mastergeorge, A. M., & Pettygrove, S. (2018). Characterizing health disparities in the age of autism diagnosis in a study of 8-year-old children. *Journal of Autism and Developmental Disorders, 48*(7), 2396–2407.

Paynter, J., Trembath, D., & Lane, A. (2018). Differential outcome subgroups in children with autism spectrum disorder attending early intervention. *Journal of Intellectual Disability Research, 62*(7), 650–659.

Pichitpunpong, C., Thongkorn, S., Kanlayaprasit, S., Yuwattana, W., Plaingam, W., Sangsuthum, S., … Sarachana, T. (2019). Phenotypic subgrouping and multi-omics analyses reveal reduced diazepam-binding inhibitor (DBI) protein levels in autism spectrum disorder with severe language impairment. *PLoS One, 14*(3).

Pickles, A., Anderson, D. K., & Lord, C. (2014). Heterogeneity and plasticity in the development of language: A 17-year follow-up of children referred early for possible autism. *Journal of Child Psychology and Psychiatry, 55*(12), 1354–1362.

Pickles, A., McCauley, J. B., Pepa, L. A., Huerta, M., & Lord, C. (2020). The adult outcome of children referred for autism: Typology and prediction from childhood. *Journal of Child Psychology and Psychiatry, 61*(7), 760–767.

Piven, J., & Rabins, P. (2011). Autism spectrum disorders in older adults: Toward defining a research agenda. *Journal of the American Geriatrics Society, 59*(11), 2151–2155.

Pohl, A., Cassidy, S., Auyeung, B., & Baron-Cohen, S. (2014). Uncovering steroidopathy in women with autism: A latent class analysis. *Molecular Autism, 5*(1), 27.

Pry, R., Bodet, J., Pernon, E., Aussilloux, C., & Baghdadli, A. (2007). Initial characteristics of psychological development and evolution of the young autistic child. *Journal of Autism and Developmental Disorders, 37*(2), 341–353.

Qian, X. (2018). Differences in teachers verbal responsiveness to groups of children with ASD who vary in cognitive and language abilities. *Journal of Intellectual Disability Research, 62*(6), 557–568.

Rapin, I., Dunn, M. A., Allen, D. A., Stevens, M. C., & Fein, D. (2009). Subtypes of language disorders in school-age children with autism. *Developmental neuropsychology, 34*(1), 66–84.

Ring, H., Woodbury-Smith, M., Watson, P., Wheelwright, S., & Baron-Cohen, S. (2008). Clinical heterogeneity among people with high functioning autism spectrum conditions: Evidence favouring a continuous severity gradient. *Behavioral and Brain Functions, 4*(1), 11.

Rødgaard, E. M., Jensen, K., Vergnes, J. N., Soulières, I., & Mottron, L. (2019). Temporal changes in effect sizes of studies comparing individuals with and without autism: A meta-analysis. *JAMA Psychiatry, 76*(11), 1124–1132.

Rommelse, N. N., van der Meer, J. M., Hartman, C. A., & Buitelaar, J. K. (2016). Cognitive profiling useful for unraveling cross-disorder mechanisms: Support for a step-function endophenotype model. *Clinical Psychological Science, 4*(6), 957–970.

van Rooden, S. M., Heiser, W. J., Kok, J. N., Verbaan, D., van Hilten, J. J., & Marinus, J. (2010). The identification of parkinson's disease subtypes using cluster analysis: A systematic review. *Movement Disorders, 25*(8), 969–978.

Ros, R., & Graziano, P. A. (2019). A transdiagnostic examination of self-regulation: Comparisons across preschoolers with ASD, ADHD, and typically developing children. *Journal of Clinical Child & Adolescent Psychology*, 1–16.

Rubenstein, E., Wiggins, L. D., Schieve, L. A., Bradley, C., DiGuiseppi, C., Moody, E., … Pence, B. W. (2019). Associations between parental broader autism phenotype and child autism spectrum disorder phenotype in the study to explore early development. *Autism, 23*(2), 436–448.

Sacco, R., Lenti, C., Saccani, M., Curatolo, P., Manzi, B., Bravaccio, C., & Persico, A. M. (2012). Cluster analysis of autistic patients based on principal pathogenetic components. *Autism Research, 5*(2), 137–147.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal, 8*(1), 289.

Seynhaeve, I., & Nader-Grosbois, N. (2008). Sensorimotor development and dysregulation of activity in young children with autism and with intellectual disabilities. *Research in Autism Spectrum Disorders, 2*(1), 46–59.

Shen, J. J., Lee, P. H., Holden, J. J., & Shatkay, H. (2007). Using cluster ensemble and validation to identify subtypes of pervasive developmental disorders. *AMIA Annual Symposium Proceedings, 2007*, 666–670.

Shogren, K. A., Shaw, L. A., Wehmeyer, M. L., Thompson, J. R., Lang, K. M., Tassé, M. J., & Schalock, R. L. (2017). The support needs of children with intellectual disability and autism: Implications for supports planning and subgroup classification. *Journal of Autism and Developmental Disorders, 47*(3), 865–877.

Silleresi, S., Prévost, P., Zebib, R., Bonnet-Brilhault, F., Conte, D., & Tuller, L. (2020). Identifying language and cognitive profiles in children with ASD via a cluster analysis exploration: Implications for the new ICD-11. *Autism Research, 00*, 1–13.

Simpson, K., Adams, D., Alston-Knox, C., Heussler, H. S., & Keen, D. (2019). Exploring the sensory profiles of children on the autism spectrum using the short sensory profile-2 (SSP-2). *Journal of Autism and Developmental Disorders, 49*(5), 2069–2079.

Smith, V., Mirenda, P., & Zaidman-Zait, A. (2007). Predictors of expressive vocabulary growth in children with autism. *Journal of Speech, Language, and Hearing Research, 50*, 149–160.

Solari, E. J., Grimm, R. P., McIntyre, N. S., Zajic, M., & Mundy, P. C. (2019). Longitudinal stability of reading profiles in individuals with higher functioning autism. *Autism, 23*(8), 1911–1926.

Solomon, M., Iosif, A.-M., Reinhardt, V. P., Libero, L. E., Nordahl, C. W., Ozonoff, S., … Amaral, D. G. (2018). What will my child's future hold? Phenotypes of intellectual development in 2–8-year-olds with autism spectrum disorder. *Autism Research, 11*(1), 121–132.

Spiker, D., Lotspeich, L. J., Dimiceli, S., Myers, R. M., & Risch, N. (2002). Behavioral phenotypic variation in autism multiplex families: Evidence for a continuous severity gradient. *American Journal of Medical Genetics, 114*(2), 129–136.

Spikol, A., McAteer, D., & Murphy, J. (2019). Recognising autism: A latent transition analysis of parental reports of child autistic spectrum disorder "red flag" traits before and after age 3. *Social Psychiatry and Psychiatric Epidemiology, 54*(6), 703–713.

Stevens, E., Dixon, D. R., Novack, M. N., Granpeesheh, D., Smith, T., & Linstead, E. (2019). Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *International Journal of Medical Informatics, 129*, 29–36.

Storlie, C. B., Myers, S. M., Katusic, S. K., Weaver, A. L., Voigt, R. G., Croarkin, P. E., … Port, J. D. (2018). Clustering and variable selection in the presence of mixed variable types and missing data. *Statistics in Medicine, 37*(19), 2884–2899.

Sullivan, M. O., Gallagher, L., & Heron, E. A. (2019). Gaining insights into aggressive behavior in autism spectrum disorder using latent profile analysis. *Journal of Autism and Developmental Disorders, 49*(10), 4209–4218.

Syriopoulou-Delli, C. K., & Papaefstathiou, E. (2020). Review of cluster analysis of phenotypic data in autism spectrum disorders: Distinct subtypes or a severity gradient model? *International Journal of Developmental Disabilities, 66*(1), 13–21.

Tanaka, S., Oi, M., Fujino, H., Kikuchi, M., Yoshimura, Y., Miura, Y., … Ohoka, H. (2017). Characteristics of communication among Japanese children with autism spectrum disorder: A cluster analysis using the children's communication checklist-2. *Clinical Linguistics & Phonetics, 31*(3), 234–249.

Taylor, S., Asmundson, G. J., & Carleton, R. N. (2006). Simple versus complex PTSD: A cluster analytic investigation. *Journal of Anxiety Disorders, 20*(4), 459–472.

Tomaszewski, B., Smith DaWalt, L., & Odom, S. L. (2019). Growth mixture models of adaptive behavior in adolescents with autism spectrum disorder. *Autism, 23*(6), 1472–1484.

Tomchek, S. D., Little, L. M., Myers, J., & Dunn, W. (2018). Sensory subtypes in preschool aged children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 48*(6), 2139–2147.

Trantou, A., Carlsen, H. K., Anderson, C., & Steingrimsson, S. (2021). Sickness absence Recommendation among outpatients with ADHD and comorbidity: A latent class analysis. *Journal of Attention Disorders, 25*(2), 209–216, 1087054718780338.

Uljarević, M., Frazier, T. W., Phillips, J. M., Jo, B., Littlefield, S., & Hardan, A. Y. (2020). Quantifying research domain criteria social communication subconstructs using the social communication questionnaire in youth. *Journal of Clinical Child & Adolescent Psychology*, 1–11.

Uljarević, M., Lane, A., Kelly, A., & Leekam, S. (2016). Sensory subtypes and anxiety in older children and adolescents with autism spectrum disorder. *Autism Research, 9* (10), 1073–1078.

Uljarević, M., Phillips, J. M., Schuck, R. K., Schapp, S., Solomon, E. M., Salzman, E., … Hardan, A. Y. (2020). Exploring social subtypes in autism spectrum disorder: A preliminary study. *Autism Research, 13*(8), 1335–1342.

Vaidya, C. J., You, X., Mostofsky, S., Pereira, F., Berl, M. M., & Kenworthy, L. (2020). Data-driven identification of subtypes of executive function across typical development, attention deficit hyperactivity disorder, and autism spectrum disorders. *Journal of Child Psychology and Psychiatry, 61*(1), 51–61.

Vargason, T., Frye, R. E., McGuinness, D. L., & Hahn, J. (2019). Clustering of co-occurring conditions in autism spectrum disorder during early childhood: A retrospective analysis of medical claims data. *Autism Research, 12*(8), 1272–1285.

Veatch, O., Veenstra-VanderWeele, J., Potter, M., Pericak-Vance, M. A., & Haines, J. (2014). Genetically meaningful phenotypic subgroups in autism spectrum disorders. *Genes, Brain and Behavior, 13*(3), 276–285.

Venker, C. E., Ray-Subramanian, C. E., Bolt, D. M., & Weismer, S. E. (2014). Trajectories of autism severity in early childhood. *Journal of Autism and Developmental Disorders, 44*(3), 546–563.

Verté, S., Geurts, H. M., Roeyers, H., Rosseel, Y., Oosterlaan, J., & Sergeant, J. A. (2006). Can the children's communication checklist differentiate autism spectrum subtypes? *Autism, 10*(3), 266–287.

Visser, J. C., Rommelse, N. N., Lappenschaar, M., Servatius-Oosterling, I. J., Greven, C. U., & Buitelaar, J. K. (2017). Variation in the early trajectories of autism symptoms is related to the development of language, cognition, and behavior problems. *Journal of the American Academy of Child & Adolescent Psychiatry, 56*(8), 659–668.

Voorspoels, W., Rutten, I., Bartlema, A., Tuerlinckx, F., & Vanpaemel, W. (2018). Sensitivity to the prototype in children with high-functioning autism spectrum disorder: An example of bayesian cognitive psychometrics. *Psychonomic Bulletin & Review, 25*(1), 271–285.

Waddington, F., Hartman, C., de Bruijn, Y., Lappenschaar, M., Oerlemans, A., Buitelaar, J., … Rommelse, N. (2018). An emotion recognition subtyping approach to studying the heterogeneity and comorbidity of autism spectrum disorders and attention-deficit/hyperactivity disorder. *Journal of Neurodevelopmental Disorders, 10* (1), 31.

Walker, S. J., Langefeld, C. D., Zimmerman, K., Schwartz, M. Z., & Krigsman, A. (2019). A molecular biomarker for prediction of clinical outcome in children with ASD, constipation, and intestinal inflammation. *Scientific Reports, 9*(1), 1–13.

Wei, X., Christiano, E. R., Yu, J. W., Wagner, M., & Spiker, D. (2015). Reading and math achievement profiles and longitudinal growth trajectories of children with an autism spectrum disorder. *Autism, 19*(2), 200–210.

Widiger, T. A. (1992). Categorical versus dimensional classification: Implications from and for research. *Journal of Personality Disorders, 6*(4), 287–300.

Wiggins, L. D., Robins, D. L., Adamson, L. B., Bakeman, R., & Henrich, C. C. (2012). Support for a dimensional view of autism spectrum disorders in toddlers. *Journal of Autism and Developmental Disorders, 42*(2), 191–200.

Wiggins, L. D., Rubenstein, E., Daniels, J., DiGuiseppi, C., Yeargin-Allsopp, M., Schieve, L. A., … Reyes, N. (2019). A phenotype of childhood autism is associated with preexisting maternal anxiety and depression. *Journal of Abnormal Child Psychology, 47*(4), 731–740.

Wiggins, L. D., Tian, L. H., Levy, S. E., Rice, C., Lee, L. C., Schieve, L., … Landa, R. (2017). Homogeneous subgroups of young children with autism improve phenotypic characterization in the study to explore early development. *Journal of Autism and Developmental Disorders, 47*(11), 3634–3645.

Wildes, J. E., & Marcus, M. D. (2013). Alternative methods of classifying eating disorders: Models incorporating comorbid psychopathology and associated features. *Clinical Psychology Review, 33*(3), 383–394.

Wing, L., & Potter, D. (2002). The epidemiology of autistic spectrum disorders: Is the prevalence rising? *Mental Retardation and Developmental Disabilities Research Reviews, 8*(3), 151–161.

Wolfers, T., Floris, D. L., Dinga, R., van Rooij, D., Isakoglou, C., Kia, S. M., … Peng, H. (2019). From pattern classification to stratification: Towards conceptualizing the heterogeneity of autism spectrum disorder. *Neuroscience & Biobehavioral Reviews, 104*, 240–254.

Wu, Y. P., Aylward, B. S., Roberts, M. C., & Evans, S. C. (2012). Searching the scientific literature: Implications for quantitative and qualitative reviews. *Clinical Psychology Review, 32*(6), 553–557.

Zaidman-Zait, A., Mirenda, P., Szatmari, P., Duku, E., Smith, I. M., Zwaigenbaum, L., … Bennett, T. (2020). Profiles and predictors of academic and social school functioning among children with autism spectrum disorder. *Journal of Clinical Child & Adolescent Psychology*, 1–13.

Zhao, X., Rangaprakash, D., Yuan, B., Denney, T. S., Jr., Katz, J. S., Dretsch, M. N., & Deshpande, G. (2018). Investigating the correspondence of clinical diagnostic grouping with underlying neurobiological and phenotypic clusters using unsupervised machine learning. *Frontiers in Applied Mathematics and Statistics, 4*, 25.

Zheng, S., Hume, K. A., Able, H., Bishop, S. L., & Boyd, B. A. (2020). Exploring developmental and behavioral heterogeneity among preschoolers with ASD: A cluster analysis on principal components. *Autism Research, 13*(5), 796–809.

Joost Agelink van Rentergem is a postdoctoral researcher at the University of Amsterdam. Marie Deserno is a postdoctoral fellow at the Max Planck Institute for Human Development in Berlin. Hilde Geurts is a professor in clinical neuropsychology and professor by special appointment (focus Autism: Cognition across the life span) at the University of Amsterdam.